THEORETICAL ANALYSES OF CODON USAGE PATTERNS

IN DNA.

FRANCIS GEORGE WRIGHT
PH.D.
UNIVERSITY OF EDINBURGH
1987

I declare that the work presented here is original and my own.

Frank Wright

30th December 1987

Abstract.

The growing body of DNA sequence data has revealed that the majority of genes show unequal usage of the alternative codons for each amino acid. Three theoretical analyses of codon usage patterns were undertaken: an exploration of patterns of codon usage, an investigation of G+C content and codon usage in human genes, and the development of a simple measure of bias in synonymous codon usage.

Codon usage data from a wide range of genomes and genes were analysed using correspondence analysis, a multivariate data reduction method, to extract the main features present. Over 40 per cent of the codon usage variation was displayed using two 2D plots to display the first four dimensions produced by the correspondence analysis.

The first dimension accounted for about one-fifth of the codon usage variation of the 428 genes studied. This dimension appeared to be very similar to third position G+C content confirming the latter as the most important factor influencing codon usage patterns. Intra-specific codon usage variation in multicellular organisms and inter-specific codon usage variation in unicellular organisms were well explained by G+C variation at synonymous sites. Intra-specific codon usage variation in unicellular organisms was almost independent of G+C content. The intra-specific codon usage patterns of yeast and *E.coli* showed considerable variation that confirmed the known bias of highly expressed genes in these two species. Distantly related bacterial species like *E.coli* and *B.subtilis* appeared more similar in codon usage than *E.coli* and yeast. The known DNA base compositional bias of animal mitochondrial genomes accounted for a large proportion of the overall codon usage variation.

The study of G+C content and codon usage in 135 human genes revealed that there are correlations between the G+C content of each of the three codon positions thus suggesting that G+C content in nonsynonymous and synonymous sites are correlated. These results are consistent with a model of human codon usage where each gene is subject to mutation pressure to alter G+C content. However, such mutational changes are subject to selective

constraints and only conservative amino-acid replacements appear to be tolerated.

A simple unbiased estimator of synonymous codon usage bias, $\hat{N}_c^p$, has been developed based on the effective number of alleles concept used in population genetics. $\hat{N}_c^p$ is a distance measure from equal usage of synonymous codons and has a range from 20 (total bias i.e. only one codon used in each amino-acid) to 61 (no bias).

Appropriate statistical methods for the analysis of codon usage patterns are briefly discussed in the final chapter, along with recent work on models based on population genetics theory.

Acknowledgements.

It is not possible to mention here everyone who contributed to the work contained in this thesis. I would however like to express my gratitude to the many people in, or connected with, the Genetics Department at Edinburgh for their help.

I would like to express special thanks to the following: my initial supervisor, Alan Robertson, for drawing my attention to the phenomenon of codon usage bias; Susan Brotherstone, for computing advice; Paul Sharp, for many conversations about codon usage; and above all, Bill Hill for his supervision in the latter stages of this work, and in particular for inspiring me to play with algebra again.

The first three years of this work were made possible by a SERC studentship; the latter part by the provision by my current employers, the Scottish Office, of an undemanding 8 to 4 job and of some special leave. I would like to genuinely thank my colleagues at the Scottish Office Computer Service for their understanding of my spare time pursuit. I would also like to thank the Science Faculty for granting an extension to allow this thesis to be completed.

Finally, I must thank/apologise-to anyone who has had to endure my anti-social lifestyle of late (i.e. the last few years) and still kept faith in me; especially my wife Anne, who was under the impression that she married a final-year Ph.D student four years ago today.

Frank Wright
23rd December 1987

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION.

This chapter reviews the broad area of codon usage bias as it was at the end of 1985. The three results chapters (2,3,4) contain more detailed and more up-to-date reviews of the literature.

1.1. Codon Usage Bias.

The study of molecular evolution has entered a new era. The recent breakthroughs in molecular genetical techniques have already resulted in large quantities of detailed information on genomic structure. Our understanding of genetic organisation has been transformed. Gene families and movable genetic elements are now known to be common. Pseudogenes and introns have been discovered. Entire genomes have been sequenced e.g. the 172,282bp Epstein-Barr Virus genome (Baer *et al.* 1984). Sequence data is now available from a wide range range of genes, species and genome types (Nuclear eukaryotic, Prokaryotic, Mitochondrial, Viral, Phage, and Chloroplast). Up-to-date DNA sequence libraries are maintained by EMBL (Europe) and Genbank (USA). The new DNA technology has allowed novel approaches to the study of gene expression, mutagenesis, DNA repair systems, molecular function and other areas of molecular biology. Detailed analysis of DNA sequence data allows unprecedented insights into evolution at the nucleotide level.

The genetic code, proposed in 1966, has been confirmed (although slight deviations from the universal code do exist). All codons are indeed used, although a wealth of literature (see Ikemura 1985a) has concluded that codons specifying the same amino acid are not used equally. Such a finding was unexpected: synonymous mutations were thought likely to be selectively neutral (Kimura 1968; King & Jukes 1969). The existence of synonymous codon usage (SCU) bias has been taken as evidence of selection acting below the level of protein structure (e.g. Grantham (1980a), Modiano *et al.* (1981); Ikemura (1981a,b;1982). Estimates of the rate of nucleotide substitution at the third codon position of mammalian globin genes (Li *et al.* 1981) is lower than

that for pseudogenes, suggesting the possible action of weak selection.

Such bias in the usage of synonymous codons appears to be a common phenomenon: the particular type of bias appears to be dependent on genome type (Grantham *et al.* 1980a) to a large extent. There is, however, variation in codon usage among genes from the same genome. For example in yeast and *E.coli*, the most biased genes are those that are highly expressed (Grantham *et al.* 1981; Bennetzen & Hall 1982), whereas in vertebrates SCU variation appears to be largely explained by G+C variation in the third codon position (Bernardi & Bernardi 1985). The observed degree of SCU bias can be extreme. As an illustration, the codon usage patterns of a selection of genes are shown in Table 1.1.

Various hypotheses have been put forward to explain the phenomenon of SCU bias (see Ikemura (1985a) for a recent review). The three most important factors appear to be G+C content, tRNA abundance and codon-anticodon interaction. However these factors do not apply equally to all genomes. In particular, there appears to be a clear difference between unicellular organisms and multicellular organisms w.r.t. codon usage patterns.

It is generally assumed that there is no relationship between SCU bias and amino acid usage. Hence codon usage bias and SCU bias usually refer to the same phenomenon. However, no assumption regarding the relationship between amino acid usage and synonymous codon usage is made at the onset of this research.

Table 1.1: Codon Usage for a selection of genes.

		E.coli		yeast		human	
		lpp	lacI	g3pdh	trp5	aglo	albu
Phe	TTT	0	3	0	17	0	25
	TTC	0	1	11	15	7	10
Leu	TTA	0	5	0	15	0	10
	TTG	0	5	20	24	0	13
Leu	CTT	0	2	0	4	1	19
	CTC	0	3	0	4	2	7
	CTA	0	0	1	11	1	3
	CTG	9	25	0	4	14	12
Ile	ATT	0	10	7	19	0	4
	ATC	2	7	12	12	0	4
	ATA	0	1	0	3	0	1
Met	ATG	2	9	7	10	2	6
Val	GTT	3	8	23	24	1	12
	GTC	0	9	12	24	3	7
	GTA	2	3	0	6	0	8
	GTG	1	14	0	7	9	16
Ser	TCT	3	7	11	21	3	3
	TCC	2	6	14	8	4	7
	TCA	0	4	0	6	0	6
	TCG	0	5	0	1	0	3
Pro	CCT	0	0	1	11	2	10
	CCC	0	6	0	4	3	6
	CCA	0	2	10	15	0	7
	CCG	0	6	0	1	2	1
Thr	ACT	4	3	10	25	0	7
	ACC	0	11	13	10	9	9
	ACA	0	1	0	9	0	11
	ACG	0	4	0	2	0	2
Ala	GCT	8	3	23	37	3	31
	GCC	0	14	10	14	11	14
	GCA	3	5	0	12	0	16
	GCG	1	22	0	1	7	2

Table 1.1 contd.: Codon Usage for a selection of genes.

		E.coli		yeast		human	
		lpp	lacI	g3pdh	trp5	aglo	albu
Tyr	TAT	0	5	0	9	1	13
	TAC	1	3	10	10	2	6
Ter	TAA						
	TAG						
His	CAT	0	3	0	11	0	11
	CAC	0	4	8	8	10	5
Gln	CAA	0	13	6	21	1	10
	CAG	5	15	0	6	0	9
Asn	AAT	0	7	0	11	0	11
	AAC	6	5	14	15	4	6
Lys	AAA	6	10	2	18	1	40
	AAG	1	1	24	26	10	20
Asp	GAT	2	10	7	26	0	25
	GAC	6	7	13	14	8	11
Glu	GAA	0	10	15	33	0	39
	GAG	0	5	0	9	4	23
Cys	TGT	0	2	2	6	0	15
	TGC	1	1	0	1	1	20
Ter	TGA						
Trp	TGG	0	2	3	5	1	2
Arg	CGT	3	2	0	3	1	3
	CGC	1	10	0	1	0	1
	CGA	0	4	0	0	0	3
	CGG	0	2	0	0	1	2
Ser	AGT	0	5	0	6	0	6
	AGC	2	6	0	1	4	3
Arg	AGA	0	1	11	22	0	13
	AGG	0	0	0	2	1	5
Gly	GGT	2	6	24	52	2	3
	GGC	1	11	0	7	5	3
	GGA	0	1	0	3	0	6
	GGG	0	4	0	4	0	2
L_c		77	359	331	708	141	608
CBI		0.84	0.18	0.99	0.45	n/a	n/a

Table 1.1 contd.: Codon Usage for a selection of genes.

Note:

(a) aglo = human α-globin; albu = human albumin.

(b) lpp (*E.coli*) and g3pdh (yeast) are highly expressed genes, unlike lacI (*E.coli*) and trp5 (yeast). The two human genes differ in the G+C content of the third codon position: α-globin (89%), albumin (39%).

(c) L_c = gene length in codons, excluding start and stop codons.

(d) CBI = codon bias index of Bennetzen & Hall (1982). see text for details.

1.2. Patterns of Codon Usage.

The first extensive analysis of codon usage data was carried out by Grantham's group (Grantham *et al.* 1980a,b;1981). They noted that all genes within a genome tend to have a similar codon usage. This was deduced from the discrete clustering obtained by using cluster analysis along with correspondence analysis. Highly and lowly expressed bacterial genes (mainly *E.coli*) also formed separate clusters. Another important, but less quoted, finding was that the most important single factor in explaining the inter- and intra- specific variation in codon usage in the set of genes studied was third position G+C content. The second most important factor was the relative usage of A and T in the third codon position.

A detailed discussion of codon usage patterns from a wide range of species and genome types is given in the following sections. Base composition differences account for a proportion of both inter- and intra-specific SCU differences. This factor is discussed first. The observation that the SCU pattern of an individual gene tends to conform to that of the species average implies that separate studies of codon usage for each species would be a useful approach. However, closely related species (e.g. the two enteric bacteria, *E.coli* and *S.typhimurium*) have very similar SCU patterns. With this in mind, codon usage patterns are discussed in terms of broad genome types.

1.2.1. DNA Base Composition.

Estimates of the mean G+C content of an organisms' DNA first became available about 30 years ago, along with rough estimates of the heterogeneity based on fragmented stretches of DNA. These studies were reviewed by Sueoka (1961a), Hill (1966), and Storck & Alexopoulos (1970). The mean G+C content of unicellular organisms (bacteria, protozoa, fungi, and algae) can vary over a wide range (from 25% to 75% G+C), but there is little G+C variation between DNA fragments drawn from one organism. Multicellular organisms (higher plants, invertebrates, vertebrates) have very similar mean G+C values (approximately 40% to 42%) but show extensive within species G+C heterogeneity. Further discussion of this literature is contained in chapter 3.

Mitochondrial and chloroplast genomes tend in general to be low in G+C content (see section 1.2.2.3), although vertebrate mtDNA have relatively high G+C values. Invertebrate mitochondrial genomes have a range of mean G+C content from 21% to 43%, whereas those of vertebrates show a range from 37% to 50% (Brown 1983). The complementary strands of vertebrate mitochondrial genomes also differ in G+T content, resulting in a heavy and a light strand. The codon usage patterns of bacteriophage and eukaryotic viruses have been reviewed by Grantham *et al.* (1985). There exists considerable inter- and intra-specific variation in the base composition of the third codon position.

Not all DNA, especially in higher eukaryotes, codes for proteins. Therefore, the above estimates relating to mean G+C content and G+C heterogeneity estimates for multicellular organisms must be interpreted with some caution when making inferences about coding DNA. However, it is now known that vertebrate coding DNA is about 10% higher in G+C content than the surrounding DNA and that nearly all the differences in G+C content between vertebrate genes are due to G+C differences at the third codon position (Ikemura 1985a; Bernardi *et al.* 1985).

Differences in G+C content between bacterial species are partly reflected in differences in amino acid composition (Sueoka 1961b, Elton 1973, Clarke 1983, Muto *et al.* 1984). However, most inter-specific G+C differences are due to all three codon positions (Bibb *et al.* 1984). The importance of the third

codon position has been confirmed by an analysis of 302 genes from a wide range of (49) genomes (Bernardi & Bernardi 1985): a strong positive correlation exists between third position G+C content and the total G+C of a gene.

1.2.2. Genomic Patterns.

The original "Genome Hypothesis" (Grantham *et al.* 1980a) stated that each gene tended to follow the average codon usage pattern of the species. Species that are closely related taxonomically (e.g. vertebrate species, or enteric bacterial species) have similar patterns of codon usage. Ikemura (1985a) has suggested that unicellular and multicellular organisms represent distinct groups w.r.t. codon usage patterns. This distinction is based on the nature of tRNA gene expression in the cells of unicellular and multicellular organisms (see section 1.3.2). Two other distinct groups exist: organisms that rely on the tRNA pools of another organisms (i.e. bacteriophages/ eukaryotic viruses) and organelle genomes. Codon usage patterns will now be discussed within this framework.

1.2.2.1. Unicellular Genomes.

The two most studied organisms w.r.t. codon usage patterns, *Saccharomyces cerevisiae* and *Escherichia coli*, are unicellular. Another notable feature is that neither have an extreme G+C content: yeast (42%); *E.coli* (52%). The intermediate mean G+C content and low heterogeneity of G+C content explain the fact that DNA base composition does not appear to be an important factor in codon usage patterns for these two species. However, unicellular species with extreme mean genomic G+C content do have appropriately biased codon usage patterns (e.g. bacteria, Ikemura 1985b; *Neurospora crassa*, Ikemura 1985a).

Codon usage patterns reflect taxonomic relationships. Thus *E.coli* and *Salmonella typhimurium* have similar codon usage patterns (Ikemura 1985a,b). *Bacillus subtilis* codon usage is, however, somewhat different (Ogasawara *et al.* 1985) from that of these two enteric bacteria. Considerable intra-specific variation in codon usage patterns exists in unicellular organisms. In yeast and *E.coli*, and possibly all unicellular species, there is a strong positive correlation

between the degree of SCU bias and the level of expression of the particular gene (Grantham *et al.* 1981, Bennetzen & Hall 1982, Gouy & Gautier 1982). Highly-expressed genes tend to use a subset of codons.

Bennetzen & Hall (1982) noted that highly expressed yeast genes tended to use preferentially a subset of 25 codons; *E.coli* highly-expressed genes tended to use a 22 codon subset. These preferred codons are shown in Table 1.2. The authors also devised a simple measure of codon usage bias, the "codon bias index" based on the usage of these preferred codons by a particular gene. A gene using only the preferred codons would have a value of unity. Random codon usage would result in a value of zero. Codon bias index values are given for the *E.coli* and yeast genes in Table 1.1. Further discussion of measures of degree of codon usage bias is contained in section 1.2.3.

Table 1.2: Preferred Codons in Highly Expressed Genes.

	UUU/Phe	e y	UCU/Ser		UAU/Tyr	Y	UGU/Cys
E Y	UUC/Phe	e y	UCC/Ser	E Y	UAC/Tyr		UGC/Cys
	UUA/Leu		UCA/Ser		UAA/***		UGA/***
Y	UUG/Leu		UCG/Ser		UAG/***	- -	UGG/Trp
	CUU/Leu		CCU/Pro		CAU/His	E	CGU/Arg
	CUC/Leu		CCC/Pro	E Y	CAC/His		CGC/Arg
	CUA/Leu	Y	CCA/Pro	Y	CAA/Gln		CGA/Arg
E	CUG/Leu	E	CCG/Pro	E	CAG/Gln		CGG/Arg
y	AUU/Ile	e y	ACU/Thr		AAU/Asn		AGU/Ser
E y	AUC/Ile	e y	ACC/Thr	E Y	AAC/Asn		AGC/Ser
	AUA/Ile		ACA/Thr	E	AAA/Lys	Y	AGA/Arg
- -	AUG/Met		ACG/Thr	Y	AAG/Lys		AGG/Arg
e y	GUU/Val	y	GCU/Ala		GAU/Asp	e Y	GGU/Gly
y	GUC/Val	y	GCC/Ala	E Y	GAC/Asp	e	GGC/Gly
e	GUA/Val		GCA/Ala	E Y	GAA/Glu		GGA/Gly
	GUG/Val		GCG/Ala		GAG/Glu		GGG/Gly

- Note: *E.coli* codons are marked (E) if there is only one preferred codon per amino acid, or (e) if more than one. If there are no preferred codons then there is no symbol. Yeast preferred codons are marked (Y) and (y) in a similar way. Including the two codons for Met and Trp yields Bennetzen & Hall (1982)'s results of 25 yeast and 22 *E.coli* preferred codons.

1.2.2.2. Multicellular Genomes.

The most studied multicellular taxonomic group are the vertebrates. Studies have confirmed that variation in third position G+C content is the major source of SCU bias in vertebrates (Ikemura 1985a; Bernardi *et al.* 1985). For example, human genes show a range of third position G+C content from 35% (blood Christmas factor-IX gene) to 95% (zeta-globin) (Ikemura 1985a). The G+C content of vertebrate genes is highly correlated with that of the surrounding region (Ikemura 1985a; Bernardi *et al.* 1985). The G+C content of exons is typically about 10% above that of introns and flanking DNA. Vertebrate genes appear to be imbedded in regions of DNA with a fairly constant G+C content. There is evidence that the homogeneity of this background DNA may extend for a large distance; much greater than 8 kb (Ikemura 1985a). Bernardi *et al.* (1985) have suggested that the genome of warm-blooded vertebrates may be composed of a mosaic of relatively G+C-homogeneous chunks of considerable size (>> 200 kb). These chunks may be the cause of the banding patterns of stained chromosomes, and may therefore be around 1250 kb in length.

Attempts to find tissue-specific SCU patterns initially failed (Hastings. & Emerson 1983). However, "housekeeping" genes appear to have a high G+C content (Goldman *et al.* 1984).

1.2.2.3. Organelle Genomes.

Reviews of the evolution of animal mtDNA (Brown 1985), plant and algal mtDNA and cpDNA (Palmer 1985), and of fungal mtDNA (Gillham *et al.* 1985) have been published.

Nearly all the organelle genes sequenced have been from mitochondrial genomes. The complete mitochondrial genomes of three mammal species (man, cow, and mouse) and of *Drosophila yakuba* have been sequenced (see references in Brown 1985). In general, mammal, *Drosophila*, and fungal mtDNA genes are well-represented in the EMBL and GenBank sequence libraries. In comparison, sequence data for plant and algal cpDNA is scarce, and that for plant and algal mtDNA very scarce.

The mitochondrial genomes of mammals, *Drosophila*, fungi, and possibly plants (see Palmer 1985) use codes that differ slightly from the "universal" genetic code, and from each other. Chloroplast genomes appear to conform to the "universal" code. Mitochondrial genomes have a small but complete set of tRNA genes (e.g. 22 tRNA genes in mammal and *Drosophila* mtDNA; 28 in *Neurospora crassa* mtDNA). To enable all codons to be decoded, the codon-anticodon pairings do not follow the standard wobble pairings. The original wobble hypothesis allowed for non Watson-Crick pairing between U and G, and proposed that I (Inosine - a modified A) could pair with U or C or A. This hypothesis thus allows for between two and four tRNA types per four-codon amino acid family. This means that mtDNA SCU bias is not generally explicable by the relative abundances of the tRNA molecules because there is usually only one tRNA species per amino acid. Less sequence data is available for chloroplast tRNAs, but again there appears to be a complete set. However, the existence of standard wobble pairing suggests that the number of tRNAs will be equal or greater than thirty-two (Gillman *et al.* 1985).

The size of the tRNA families in animal mtDNA, fungal mtDNA and plant cpDNA appears to be related to the size of the organelle genome. Sizes range from 16-18 kb (animal mtDNA), through 60-74 kb (fungal mtDNA) and 105 kb (broad bean mtDNA), to 130-135 kb (flowering plant cpDNA) (Fincham 1983).

Studies of the total G+C content of organelle genomes reveal that biased DNA base composition is common in many organelle genomes. Animal mtDNA genomes range from 21% to 50% in total G+C content with "warm-blooded" mtDNAs tending to also show differences in G+T content between the two complementary strands (Brown 1983). The chloroplast genomes of plants and algae exhibit a range of total G+C content from 17% to 47% G+C. Fungal mtDNA, e.g. yeast (Bonitz *et al.* 1980), also tends to have a low G+C content.

The substitution rates observed in organelle DNA has been reviewed by Brown (1985). Angiosperm cpDNA and mtDNA both appear to evolve very much slower than fungal and mammalian mtDNA. The rate for nuclear mammalian DNA lies between these estimates. There is some evidence that *Drosophila* mtDNA is evolving at about the same rate as nuclear DNA.

1.2.2.4. Bacteriophage and Viral Genomes.

A comparison of third position G+C patterns between human viruses and human nuclear genes has been carried out by Grantham *et al.* (1985). Viral genes, like the genes of their host, show a wide range of third position G+C content. However viral genes tend to have lower third position G+C values (ranging from 25% to 85%) than human genes (55% to 90%), although there is some overlap.

Grantham *et al.* (1985) also compared the codon usage patterns of eukaryotic viruses and bacteriophages: a combined analysis suggests that the prokaryote/eukaryote difference is more important than genome chemical composition (RNA or DNA; single or double stranded molecule).

The relationship between the codon usage patterns of *E.coli* genes and those of phages infecting *E.coli* has been studied by Grantham *et al.* (1985), and Sharp *et al.* (1985). Both groups note that phage gene SCU bias is related to the relative concentration of tRNA species in the host cell. The effect is not as strong as found for *E.coli* highly expressed genes, and the degree of SCU bias is similar to that of *E.coli* plasmid genes (see section 1.3.2). The degree of correlation with *E.coli* tRNA abundance appears to be related to the level of expression of the particular phage gene. Grantham *et al.* (1985) also noted that the virulent T7 phage was more adapted to *E.coli* tRNA abundance than was the temperate λ phage.

1.2.3. Degree of SCU bias.

Most authors note that codon usage patterns tend to be "non-random" or "non-uniform", without explicitly stating what constitutes an unbiased pattern of codon usage. A "reference" codon usage pattern is required to allow the development of quantitative measures of the degree of codon usage bias, and appropriate statistical tests.

The choice of a reference codon usage pattern requires a consideration of the underlying evolutionary forces. This reference pattern could be based on "unbiased" codon usage, denoted H_0, or on a previously known pattern of bias, denoted H_1. The choice of the H_0 reference pattern may be based on a

uniform usage of nucleotides in synonymous positions, or may be adjusted to reflect some measure of base composition. The relationship between amino acid usage and synonymous codon usage must also be considered.

Considerable progress has been made in this area recently (see the review in chapter 4). However, there is a lack of a simple, intuitively obvious measure of codon usage bias for genes about which there is little additional information other than the codon usage data.

1.3. Factors Affecting Codon Usage.

To explain the above patterns of codon usage bias, a number of explanations have been put forward. Important reviews of factors affecting codon usage have been published recently (Ikemura 1985a,b; Li *et al.* 1985).

The discussion of putative factors influencing codon usage can be split into four broad categories: mutation, optimization of translation, avoidance of damage to the nucleic acid, and structural constraints. These factor categories do not apply equally to the four broad genome types outlined in section 1.2.2.

1.3.1. Mutation.

Both the occurrence of cryptic patterns in DNA and the overall DNA base composition may be due, in part, to mutation. The chemical instability of bases and errors in DNA replication are corrected by the DNA repair systems. The partial success of these repair systems results in actual spontaneous mutations. The movement of genetic elements may also be an important cause of mutation. Rubin (1983) has suggested that a large fraction of spontaneous mutations and chromosomal rearrangements in *Drosophila* are due to the movement of transposable elements.

In coding DNA, nucleotide substitutions are much more frequent than insertion/deletion events and the majority of substitutions are found at synonymous sites (Kreitman 1983). However, the occurrence of substitutions is not independent of neighbouring bases (Bird 1980; Kunkel *et al.* 1981; Foster *et al.* 1982) and therefore the amino acid sequence of a gene could influence third position nucleotide composition in an indirect manner. The dinucleotide

CpG is rare in vertebrate DNA and appears to be a mutational 'hotspot' (Bird 1980), and therefore NNC codons (N = any nucleotide) would be expected to be rare immediately before a codon of the form GNN. Other detailed (i.e. at the DNA level) examples of genome type specific or species-specific mutation spectra are likely to be found due to differences in exposure to mutagens and to differences in repair systems. There is some evidence that eukaryotic DNA repair systems differ from those of prokaryotes (Lawrence 1982). It therefore seems likely that patterns in DNA due to mutation will vary between organisms.

Estimates of the rates of the twelve possible substitution types have been derived (see Li *et al.* 1985) by comparing vertebrate pseudogenes with their functional counterparts. The rates calculated suggest that vertebrate DNA would equilibrate at a level of G+C content lower than that actually observed in coding DNA, if no other forces were acting. This is in agreement with the finding that vertebrate coding regions are about 10% higher in G+C content than the surrounding DNA (Bernardi *et al.* 1985).

Mutation can influence the overall DNA base composition as well as causing cryptic patterns. Mutations occurring in the mutT+ gene in *E.coli* (Treffers *et al.* 1954) resulted in a dramatic increase in A:T to C:G transversions, thereby increasing the overall mutation rate by two orders of magnitude (Fincham 1983). The wide range in total G+C content, and low within-species G+C heterogeneity, found in unicellular organisms (see section 1.2.1) has been explained in terms of mutational pressure by Sueoka (1962). He proposes that there is a optimum mean G+C content for a species, and this is achieved by selection on the forward and backward mutation rates between A:T and G:C nucleotide pairs. This mutation pressure model can be applied to Bernardi *et al.* (1985)'s model of the vertebrate genome. The differing G+C levels observed in different genomic chunks may be due to differences in local mutation rate. However, the mechanism causing such between-chunk G+C variation is as yet unknown.

Spontaneous mutation may be the cause of significant departures from uniform usage of synonymous codons (unless the mutation pressure is itself towards equal usage). Alternatively it could act so as to obscure or remove SCU bias due to "weak" selection. The relationship between base composition

and substitution rate has not been studied to any large extent. Theoretical models to explain codon usage bias are discussed in section 1.4.

1.3.2. Optimization of Translation.

The existence of synonymous codon usage bias in many genes may be due to selection acting to optimise translational speed and/or accuracy. Most authors consider translational speed to be the more important factor (see Li *et al.* (1985) for further discussion). The high cost of protein synthesis in terms of energy and substrates has been discussed by Ikemura (1985a).

Although it is convenient to consider codon usage patterns as consisting of amino acid usage and synonymous codon usage, at the level of the translational process it is the tRNA molecule that relates the amino acid to the codon. There can be more than one unique tRNA species for a particular amino acid. For example, in *E.coli*, there are two tRNA species that decode isoleucine: $tRNA_1^{Ile}$ recognises both the AUU and AUC codons; $tRNA_2^{Ile}$ recognises the AUA codon. Translational speed can be optimised by matching the demand for tRNAs (i.e. the codon usage pattern) to the abundances of the different tRNA species. Thus tRNA abundance patterns can "explain" the observed amino acid usage patterns and part of the observed SCU pattern. Note that tRNA abundance patterns cannot explain differences between the usage of the *E.coli* isoleucine codons AUU and AUC mentioned above because they are decoded by a single tRNA species. See section 1.3.2.1 for further discussion of tRNA abundance and codon usage.

The "wobble" rules governing codon-anticodon pairing were proposed by Crick in 1966. These provide a simple prediction of pairing rules. There is now a vast literature on the peculiarities of codon-anticodon pairing (see Nishimura 1978) and a large number of sequenced tRNAs (see Singhal & Fallis 1979; Gauss & Sprinzl 1984a,b). This allows the known abundance of a tRNA species to be compared to the pooled usage of the codons which it is known to decode. The two *E.coli* Ile tRNAs noted above have different bases in the first anticodon position (which binds to the third position of a codon): $tRNA_1^{Ile}$ contains a G, and $tRNA_2^{Ile}$ a modified C, in this position. (Chemical modification of nucleotides in tRNA molecules is common (Nishimura 1978). These bind to U and C, and to A, respectively. The number of tRNA species

for a particular amino acid, and the nature of the respective anticodons, can vary between species and genome-types. However, related species (e.g. *E.coli* and *Salmonella typhimurium*, Ikemura & Ozeki 1982) have very similar tRNA abundances and anticodon sequences.

The nature of the base of the anticodon can also influence the usage of codons decoded by the same tRNA species. The particular tRNA species may show differences in affinities for, or different binding strengths with, the possible codons (see Ikemura 1985a). Thus selection could act on those codons that form "optimal" codon-anticodon bonds. For example, Grosjean & Fiers (1982) have suggested that codons of the form WWY (where W = A or U, and Y = U or C) will tend to have C in the third position. This prediction is part of a wider hypothesis that suggests that codon-anticodon pairings of intermediate stength are optimal (note that C:G bonds are stronger than A:U bonds). This leads to the prediction that the AUU ile codon will be less common than the AUC ile codon in the above example. See section 1.3.2.2 for more discussion on codon-anticodon interactions.

1.3.2.1. tRNA Abundance.

Amino acid usage and part of the synonymous codon usage may be related to the abundance of tRNA species. Codon usage data pooled over tRNAs will be referred to as tRNA usage (after Ikemura 1985a) (this pooling of information removes "within tRNA" variation). Correlations between amino acid usage and tRNA abundance have been found in certain tissues of the multicellular Bombyx mori (Garel 1976) and between tRNA usage and tRNA abundance in unicellular organisms (yeast, *E.coli*, and *S.typhimurium*, (see Ikemura 1985a). *Bombyx mori* serves as a useful model of a multicellular organism: the tRNA abundance pattern appears to adjust to the codon usage patterns of those genes being expressed in a particular tissue. This has been termed a "functional adaptation of tRNA population to codon frequency" (Chevallier & Garel 1979). In contrast, the genes of unicellular organisms are all exposed to the same tRNA abundance pattern.

Unicellular genomes.

Major progress in the understanding of yeast and *E.coli* codon usage patterns was made possible by the quantification of the levels of the individual transfer RNA species within each organism (Ikemura 1980; 1981a,b; 1982: see Ikemura (1985a) for a review). The tRNA abundance patterns for these two organisms were markedly different. However each tRNA abundance pattern was highly correlated with the pooled codon usage pattern of genes found in its own genome. The tRNA abundances of two closely related enteric bacteria, *E.coli* and *S.typhimurium*, were however found to be very similar (as was their codon usage patterns).

The SCU bias of individual genes was found to be correlated with their level of expression, highly expressed genes having the most biased codon usage patterns (Nomura *et al.* 1980). Ikemura (1985a) has shown that highly expressed genes (of enteric bacterial species) have a lower synonymous substitution rate than lowly expressed genes, suggesting that selection is acting on highly expressed genes to maintain SCU bias. Kimura (1981; 1983) has developed theoretical models for codon usage patterns constrained by tRNA availability (see section 1.4).

Multicellular genomes.

Codon usage in multicellular organisms has been reviewed by Ikemura (1985a). Only a small proportion of genes are expressed in the typical cell of a multicellular organism. The tRNA abundance may adapt to the codon usage of those genes being expressed as occurs in *Bombyx mori* tissues (see above). The larger tRNA gene families of multicellular organisms compared to unicellular organisms further supports this view (e.g. *E.coli*: about $60^{tRNA}_{\wedge}$genes; yeast: about $360^{tRNA}_{\wedge}$genes; human: about $1300^{tRNA}_{\wedge}$genes) (see Birnsteil *et al.* (1972), Guthrie & Abelson (1982), Hatlen & Attardi (1971), respectively). However, a preliminary comparison of liver and muscle genes from a range of warm-blooded species has revealed no significant differences in codon usage patterns (Hastings & Emerson 1983). If, as it seems, third position G+C content is the major factor in vertebrate codon usage patterns (Ikemura

1985a), then it is unlikely that a correlation of codon usage with tRNA abundance has been the cause. Anticodons with G in the "wobble" position bind to codons with U or C in the third position: it is therefore not possible for constraints on tRNA availability to lead to selection for U or C on codons of the form NNY (where Y = U or C).

1.3.2.2. Codon-Anticodon Interaction.

Codon-anticodon interactions have been reviewed by Li *et al.* (1985) and by Ikemura (1985a,b). Various authors have attempted to explain SCU bias, not attributable to tRNA abundance, in terms of interactions between the codon and the anticodon and bases close to the anticodon.

Three categories of hypothesis can be recognised:

I. Optimal Codon Anticodon Interaction Energy (or Pyrimidine Bias).

Grosjean & Fiers (1982) observed the following synonymous codon usage pattern in MS2 RNA bacteriophage: WWY codons (where W = A or U; Y = U or C) tended to have C in the third codon position, but SSY codons (where S = G or C) tended to have U. Since G:C bonds are stronger than U:A bonds, they proposed that codon-anticodon pairings of intermediate strength were optimal for translation. Y stands for pYrimidine; hence this type of SCU bias will be referred to as Pyrimidine Bias. The pyrimidine bias is also found in highly-expressed *E.coli* genes.

II. Simple Binding Bias.

Pyrimidine Bias cannot be explained simply in terms of the properties of the base at the first anticodon position. However, there are several bases for which experimental data on tRNA affinities for particular codons is available. Details are available in Ikemura (1985a). In terms of codon usage, these results lead to the following main predictions:

first anticodon base	binding bias
Inosine	(U = C) > A
modified U	A > G

III. RNY (or Y/R Boundary Bias).

RNY codons, where R = puRine (i.e. A or G) and Y = pYrimidine (i.e. U or C), are frequently found to account for more than 25% of codons in genes from a wide range of species (Shepherd 1981). A repeating RRY (Crick 1976) or RNY (Eigen & Schuster 1979) primordial messenger RNA would have certain optimal physiological characteristics. These considerations have led to the suggestion that the surplus of RNY codons is a remnant of the original code (Shepherd 1981) but this is unlikely: such a trace would have been removed by mutation. Pieczenik (1980), in a study of *E.coli* anticodons, noted that the 3' and 5' bases next to the anticodon tended to be an R and a Y respectively. If these adjacent bases were involved in codon-anticodon interactions then Y/R codon boundaries might be subject to natural selection to maximise the interaction.

RNY codons are confounded with other factors when codon usage data are analysed. It is only possible to test the relationship of RNY versus RNR w.r.t. synonymous codon usage for five amino acids (Ile, Val, Thr, Ala, Gly). Even then the possible effects of tRNA availability can explain the observed codon usage pattern.

1.3.3. Avoidance of Damage to the Genetic Material.

The patterns of synonymous codon usage observed may be the manifestation of a strategy to avoid damage to the genetic material. Clarke (1982) has considered the interaction of G+C content and the action of mutagenic agents, and has suggested that the susceptibilty to mutation can be altered by varying DNA base composition.

Based on limited human globin data, Modiano *et al.* (1981) have suggested that human globin genes avoid using codons that are one substitution away from a stop codon (i.e. pretermination codons, PTC). The effectiveness of this evolutionary strategy has been questioned by Kimura (1983), because the selective advantage is too low, and by Golding & Strobeck (1982), because such an effect would be very small. Several authors (Kimura (1983), Li *et al.* 1985) explain the observed SCU pattern as due to these PTC codons being decoded by rare tRNA species. An alternative explanation is that the high G+C content of globins caused bias against the predominantly A+T rich "PTC" codons; the observed total absence of PTC codons is probably due to sample size.

Most bacteria possess restriction and modification systems that attack foreign DNA. Bacteriophage DNA would appear to be under selective pressure to minimise the number of sites recognised by bacterial restriction enzymes. *Bacillus* phages ϕ1, ϕ29, and SPO1 (Kruger and Bickle 1983, and the *E.coli* phage T7 (Rosenberg *et al.* 1979; Sharp *et al.* 1985) appear to use this strategy: *E.coli* phages ϕX174, fd, and G4 do not (Adams & Rothman 1982). Sharp *et al.* (1985) note that 14 out of the 20 codons that are part of vulnerable palindromes are under-represented in pooled T7 codon usage data.

1.3.4. Structural Constraints.

The conformation of the DNA molecule may be sequence dependent. Although initial attempts are being made to establish a "sequence/structure vocabulary" (Lennon & Nussinov 1984), most of the studies are based on very simple synthetic molecules, and very little is immediately applicable to codon usage studies. Any periodicities introduced into nucleic acid sequence organisation by the differing helical twist angles of the three major DNA conformations (A, B, and Z) would be difficult to detect in codon usage data.

Genomes vary in the nature of the nucleic acid (RNA or DNA) and whether it is single (ss) or double (ds) stranded. Bacteriophage and eukaryotic viral genomes show much variation in these respects. Study of the codon usage patterns of these taxonomic groups may reveal constraints on nucleic acid structure. Eukaryotic chromosomes are structurally more complex than those of prokaryotes. In particular, the binding of histone proteins to DNA

may impose evolutionary constraints or provide protection from mutagens. The periodicity of DNA-histone complexes (nucleosomes) is however about 200 bp (Lewin 1983) and is unlikely to be detectable in codon usage data.

Li *et al.* (1985) have reviewed structural constraints on mRNA in relation to codon usage. There is little useful data available to make clear inferences on structural constraints on messenger RNA. However, there may be differences between prokaryotic and eukaryotic mRNA. The eukaryotic primary transcript must undergo substantial processing unlike the immediately functional prokaryotic mRNA (Nevins 1983). The mature eukaryotic mRNA may still contain splice-site information, thus distinguishing mRNAs that have had their introns removed. Lipman and Maizel (1982) noted differences in base order and base composition between exons from genes with/without introns.

The ability of a tRNA molecule to accurately translate a codon may depend on the nature of the nucleotides on either side of the codon. This "codon context" effect has been discussed by Bossi & Roth (1980), Bossi (1983), and Li *et al.* (1985). The mechanism of such an effect is unknown but may involve nucleotides next to the codon in interactions between the mRNA and the ribosome or the tRNA on the ribosome. The effect could also be due to interactions between tRNA molecules bound to the ribosome at the same time.

Li *et al.* (1985) has suggested that Blaisdell's (1983a,b) observation that eukaryotic genes avoid long runs of G and C, and also of A and T may be due to context effects. Wada & Suyama (1983) also suggest a similar functional constraint in bacterial and phage DNA.

1.4. Theoretical Models of Codon Usage.

The development of theoretical population genetics models of codon usage bias is at an early stage. While many authors discuss putative selective constraints acting on codons, only Ikemura (1981a,b) and Kimura (1981,1983) have attempted to produce quantitative models. Both authors consider a genome where tRNA abundance and codon usage are coevolving, although the tRNA abundance is assumed fixed to simplify the model. Any mutation at a synonymous site of a gene which decreases the correlation between tRNA

abundance and codon usage will lead to a reduction in fitness (i.e. translational efficiency and/or fidelity) and will be subject to stabilising selection.

Kimura (1981,1983) considered unicellular genomes where there is a considerable intra-specific range in SCU bias which is related to the level of expression of the particular gene. Ikemura's simple model states that the selective value of a synonymous mutation is proportional to the expression level of the gene.

Kimura (1981,1983) attempted to answer two general population genetics problems: (1) what level of selection is required to explain the observed degree of SCU bias? ; (2) how much does selection reduce the observed substitution rate compared to that expected when no selection is acting? A simplified model of synonymous codon usage is assumed, with only two types of synonymous base and with equal mutation rates between the two alternatives. Assuming the optimal codon usage pattern is different from the uniform usage produced by mutation, Kimura considers a model of stabilising selection acting on codons so as to maintain the match with the (fixed) tRNA abundance pattern. To maintain a relative SCU bias of 0.7 to 0.3 (i.e. a relative usage of 0.7 to 0.3 between the two synonymous alternatives) requires only weak selection at individual sites. The product of effective population size and the selection coefficient is approximately 0.21.

The relationship between observed mutation rate and SCU bias was also studied. With a relative SCU bias of 0.7 to 0.3 for the two synonymous alternatives, a relative mutation rate of 0.89 times that of the neutral rate is expected; for 0.9 to 0.1, the expected relative rate is 0.49. Therefore extreme SCU bias involves a halving of the observed substitution rate. These results are approximations from a simple model, but they do suggest that weak selection can account for the observed SCU bias.

Unicellular and multicellular organisms will differ in effective population size and mutation rate. The larger size of tRNA gene families in multicellular organisms may mean that tRNA abundance and codon usage are not as closely coupled as they appear to be in unicellular organisms.

1.5. Plan of Research.

A considerable amount is now known about the codon usage patterns of a few genomes. In addition, the relationship between intra- and inter-specific codon usage patterns has been displayed for 119 genes from a range of genomes (Grantham *et al.* 1980a). Grantham's group suggested that the two factors that explained most of the variation in codon usage patterns were third position G+C content and the relative use of A and T in the third position. Subsequent studies of vertebrate genes (e.g. Ikemura 1985a) have shown third position G+C variation as the major source of intra-specific SCU variation in vertebrates. However, this factor is likely to affect other genomes: the work of Sueoka (1961a) on base compositional trends between and within species suggests that G+C content will be a dominant factor influencing codon usage patterns.

Other studies have identified additional factors likely to cause variation in codon usage patterns (e.g. the expression level of unicellular genes). Factors likely to affect all genes (and thus not likely to cause variation in codon usage patterns) include the preferential use of RNY-type codons and the avoidance of pre-termination codons. Both the amount of sequence data, and the number of species and genome-types for which such data is available, has increased considerably. The opportunity therefore exists to look again at patterns of codon usage.

The data exploration methods used by Grantham's group were correspondence analysis and cluster analysis. The former method helped to extract the main features of the codon usage data studied; the latter checked for the existence of discrete groupings in the data. The main emphasis was on the finding of clusters rather than the description of factors. The proportion of the variation explained was not quantified and inter- and intra-specific variation were not distinguished. Greenacre's review of correspondence analysis (Greenacre 1984) emphasizes the use of correspondence analysis in data exploration. In particular, the proportion of the variation in the data explained by each factor can be quantified and previously identified factors (e.g. G+C content) can be plotted alongside the original data.

For three reasons: (1) the availability of a more extensive, more representative data library of DNA sequence data; (2) our increased knowledge of putative factors affecting codon usage patterns; (3) the possibility of using correspondence analysis in a more quantitative manner with more emphasis on hypothesis development, a re-examination of codon usage patterns was planned. This work is detailed in chapter 2.

In chapter 3, an aspect of the relationship between amino acid usage and synonymous codon usage is studied. The importance of G+C content in SCU variation in vertebrate genes has been noted by several authors (Ikemura 1985a, Bernardi *et al.* 1985). Codon usage data from 135 human genes compiled by Maruyama *et al.* (1986) was used to study the relationship between amino acid usage and synonymous codon usage w.r.t. G+C content. A comparison was carried out with a non-vertebrate genome, yeast.

In chapter 4, the emphasis shifts away from exploring codon usage patterns towards a more concrete task: a simple measure of SCU bias is developed so as to allow comparisons between genes. Finally, in chapter 5, the last chapter, the three separate analyses are drawn together.

CHAPTER 2
PATTERNS OF CODON USAGE

2.1. Introduction.

The first extensive analysis of codon usage patterns was carried out by Grantham and his co-workers (Grantham *et al.* 1980a,b; 1981). They employed a multivariate statistical method called correspondence analysis to extract the main features of the codon usage patterns from relatively small data sets (90, 119, and 161 genes, respectively, in the three references above). Exploratory data analysis methods like correspondence analysis are frequently used to find 'pattern' or 'structure' in the data, and thus to generate ideas for the formulation of theoretical models. While there are already several putative factors involved in determining codon usage patterns, such a preliminary data descriptive approach represents a more objective starting point than moving straight to hypothesis testing.

There are two main reasons why another Grantham-like exploratory data analysis of codon usage patterns is required. The first is due to the continuing growth in the available DNA sequence data thus offering not only a large increase over the number of genes studied by Grantham's group but also a much richer choice of species and gene types. The second reason is that a more detailed interpretation of the output of such methods is possible (Greenacre 1984) than that attempted so far in the analysis of codon usage.

Recent studies of codon usage patterns have revealed some genome-specific trends: for example, all genes in a genome tend to use a similar codon usage pattern (Grantham *et al.* 1980a,b); in *E.coli* and yeast (at least) there is a correlation between degree of codon usage bias and the level of gene expression (Ikemura 1981a, 1982). A particular emphasis in this chapter will be a search for factors that affect genes regardless of genome type.

2.2. Factors Influencing Codon Usage Patterns.

A review of factors influencing codon usage patterns was presented in the previous chapter. This section briefly reviews these in the context of an exploratory data analysis. Emphasis will be placed on which mono-nucleotides, di-nucleotides and/or tri-nucleotides are affected rather than the evolutionary forces responsible. Section 2.3.4 deals with multivariate studies of codon usage in more detail.

2.2.1. Constraints on Amino-Acid Choice.

Analyses of codon usage patterns usually disregard the contribution of amino-acid composition. While a particular amino-acid position within a protein may be under selection, the pooled nature of a codon usage table for a moderately long gene typically assumes an "average" pattern of amino-acid usage. Grantham *et al.*'s (1980b) study of the codon usage patterns of 119 genes found clear patterns in codon usage that were not apparent when amino- acid usage was studied. No significant differences in amino-acid composition have been found between highly and lowly expressed *E.coli* genes (Blake & Hinds 1984), or between mammalian genes differing in G+C content (Bernardi *et al.* 1985). However, inter-specific comparisons of the S8 and L6 ribosomal protein genes between *E.coli* and *Mycoplasma capricolum* reveal clear differences in amino-acid composition (Muto *et al.* 1984). These differences are a result of conservative amino-acid replacements that contribute to the lower G+C content of *M.capricolum* w.r.t. *E.coli*.

2.2.2. Mutation Pressure.

Sueoka (1961a) noted that bacterial and fungal species showed a wide range in their overall G+C content. A particular species appeared to have a particular level of G+C content: stretches of DNA within a genome showed much less variation than that seen between species' means. To account for this observation, Sueoka suggested a simple model of mutational pressure in which the forward and back mutation rates between A:T and G:C nucleotide pairs determined the overall G+C content of a species.

The importance of base composition has also shown up in studies of codon usage. The original correspondence analysis of Grantham's group in 1980 (Grantham *et al.* 1980a,b) displayed some of the considerable variation in codon usage between and within genomes as a two-dimensional plot. The proportion of the total variation displayed was not stated. The main factors appeared to be third position G+C content and the relative usage of U and A in the third codon position. In the analysis of 119 genes (Grantham *et al.* 1980b), seven discrete classes of gene were obtained. With respect to the first axis ("third position G+C content"), these were ordered (in decreasing G+C):

- mammals (12 genes)
- bacteria (9 genes)
- highly-expressed bacterial genes (5 genes)
- papova, adeno & hepatitus B viruses (22 genes)
- ssDNA phages & yeast and slime mould (9 genes)
- mitochondria (2 genes)

Not all genes in the analysis were put into these classes. The size of each class is given in brackets. The extreme classes at the positive end of the second axis ("third position U/A ratio") were highly expressed bacterial genes, yeast and slime mould genes plus part of the ssDNA phage and bacteria classes. The papova, adeno and hepatitus B viruses class lay at the negative end of this axis.

Nucleotide base composition differences, other than G+C content, are obvious between animal mitochondrial genomes. Brown (1983) notes that "warm- blooded" species (human, green monkey, woolly monkey, house mouse and chicken) have larger differences between the base composition of the complementary mtDNA strands than do other animal species (frog (two species), sea urchin, mussel, and *D.melanogaster*). This results in a "heavy" mtDNA strand and a "light" mtDNA strand . These results are derived from the relative ability of the strands to separate in alkaline CsCl gradients (Brown 1981). Heavy and Light strands differ in G+T content. Since most genes are

located on the heavy (G+T rich) strand, most "warm-blooded" animal mitochondrial mRNAs are rich in C+A content.

Summaries of codon usage patterns using base composition data are common. Grantham *et al.* (1985) uses a 2D plot of third position G+C content versus third position U/A as a routine method of displaying genes. Rowe *et al.* (1984) carried out a cluster analysis of nucleotide frequencies in each of the three codon positions as a method of displaying the main features of codon usage. The complexity of a codon usage pattern cannot however usually be adequately described in terms of the mono-nucleotide frequencies in each of the three codon positions, or in terms of amino-acid composition plus third position nucleotide composition. The next level of modelling is to consider factors affecting di-nucleotides. While there is abundant evidence of pattern in compilations of doublet frequencies (Nussinov 1981), this usually provides little information on the factor(s) responsible. The observed frequencies may reveal the average pattern of nonrandom mutation and/or constraints on DNA and RNA structure but it is difficult to make definite conclusions from this type of data.

An exception is the well-documented phenomenon of the under-representation of the CpG dinucleotide in eukaryotes (Bird 1980). This dinucleotide tends to mutate to CpA and TpG. This implies that codons of the type CGN and NCG will be less abundant in eukaryotes, and CAN, NCA, TGN, and NTG correspondingly more common (N represents any of the four nucleotides). Bernardi & Bernardi (1985) have noted that the ratio of CpG to GpC increases in warm-blooded vertebrate genes that have high G+C content. This suggests that CpG sites in these genes are less likely to mutate.

While it is known that nucleotide substitutions are not independent of neighbouring bases (Bird 1980), there is as yet little detailed information of other sequence-specific 'hotspots'. Such information is unlikely to be generally applicable due in part to differences in repair systems between eukaryotes and prokaryotes (Lawrence 1982).

While synonymous codon usage may mirror the actions of spontaneous mutation, another possibility is that those synonymous codons that minimise mutational damage will be subject to selection (Clarke 1970). Modiano *et al.* (1981) noted that codons that were one point mutation from a stop codon

were not used in human globin genes if there existed another synonymous choice. They suggested that selection against these "pretermination codons" might represent an evolutionary strategy for reducing the possibility of a sense codon mutating to a stop codon. Kimura (1983) has shown that the selective advantage would be very small. Golding & Strobeck (1982) have carried out a simulation that suggests that the effect on codon usage patterns would also be small.

2.2.3. Translational Efficiency and Fidelity.

Several factors influencing codon usage patterns are related to the translation of the messenger RNA into a polypeptide. A major factor in explaining intra- and inter-specific codon usage patterns, at least in unicellular organisms, is tRNA abundance. In a major review, Ikemura (1985a) lists tRNA abundance data for the closely related *E.coli* and *S.typhimurium*, and also for *S.cerevisiae*. The two bacteria have very similar tRNA pools but these are different from that of yeast. For each of these three species, the tRNA abundance is highly correlated with the codon usage patterns of highly expressed genes (e.g. ribosomal protein genes). Highly expressed genes appear to have very biased codon usage patterns to enable them to make optimal use of the tRNAs available, and thus to maximise speed of translation. Alternatively, Grosjean & Fiers (1982) have postulated that the codon usage pattern of lowly expressed genes may regulate the rate of expression. Translational fidelity may be another reason for the observed tRNA abundance patterns: errors in amino-acid incorporation increase when the concentration of the correct aminoacyl tRNA is low (Edelman & Gallant 1977).

Not all the differences between biased highly expressed genes and relatively unbiased lowly expressed genes are due to tRNA abundance. For example, there is only one tRNA species that binds to the phenylalanine codons UUU and UUC, and yet these are unequally used in both *E.coli* and yeast highly expressed genes (see Table 2.5 on page 63 for a list of most-commonly used codons for these two species).

Grosjean & Fiers (1982) have proposed that certain codons are optimal for codon-anticodon interactions. Their "Optimal Codon-Anticodon Interaction Energy" hypothesis refers to codons with a pyrimidine (i.e. U or C) in the third

position that are served by the same tRNA species. The OCAIE hypothesis states that codons of intermediate G+C content produce codon-anticodon interactions of intermediate strength. They therefore predict that codons with A and/or U in the first two codon positions and C in the third will be more common than those ending in U. Similarly, codons with G and/or C in the first two positions and U in the third will be more common than those ending in C. This third position "Pyrimidine Bias" however only offers an explanation of part of the "within tRNA" heterogeneity in the codon usage table.

Ikemura (1985a) has drawn up a list of rules to predict the relative usage of codons recognised by a single tRNA species. Several of these are based on experimental evidence: tRNAs recognising codons ending in -A and -G tend to preferentially bind to the codon ending in A; tRNAs recognising codons ending in -U, -C, and -A tend to prefer to bind to those ending in -U and -C. The OCAIE hypothesis is also included. These rules successfully predict "within tRNA" choices in Yeast and *E.coli* to a large extent.

Consideration of the identity of nucleotides on each side of the anticodons of *E.coli* tRNAs lead Pieczenik (1980) to suggest that codons of the form RNY (where R is a purine (A or G), and Y is a pyrimidine) would result in a codon-anticodon interaction over five bases rather than three, thus maximising tRNA-mRNA interaction. While there is evidence that RNY type codons are over-represented in many organisms (Shepherd 1981), this phenomena may well be due to mutation pressure and/or tRNA abundance.

2.2.4. Constraints on DNA and RNA Secondary Structure.

In the thirty years that have elapsed since the double helical nature of DNA was discovered, much progress has been made in the understanding of DNA secondary structure. However, we are only beginning to understand the relationship between sequence and structure (Lennon & Nussinov 1984).

Sharp *et al.* (1985), in a study of the total codon usage of bacteriophage T7, looked at the occurrence of potential restriction sites. Codons that could form part of short palindromes were under-represented in most cases (see Table 2.5 on page 63 for details of these codons).

Another approach is to infer structural constraints from the observed frequencies of certain oligonucleotides. Such studies within coding regions will be dominated by the protein sequence information. Blaisdell (1983a,b) has however commented on the low occurrence of long runs of weak hydrogen bonding nucleotides (U and A), and strong hydrogen bonding nucleotides (C and G), in eukaryotic genes.

2.3. Data Analysis and Codon Usage.

Given the complexity of the data, some form of data reduction is advisable. Prior to the work of Grantham's group, codon usage patterns were summarised by pooling the available data for a species or taxonomic group and indeed this practice is still common (Maruyama *et al.* 1986). This disgards intra-specific variation. Before considering the choice of a data reduction technique, it is useful to consider the nature of codon usage data.

2.3.1. The Structure of Codon Usage Data.

Codon usage data for a set of I genes can be structured in a number of ways depending on the objectives of the analysis. There are three obvious possibilities:

1. The codon usage table of each gene can be viewed in its "natural" form as a three-way contingency table made up of sixty-one sense codons and three termination codons, (assuming the "universal" genetic code), classified by their base composition in each of the three codon positions. An obvious structure for the dataset is an (I x 4 x 4 x 4) contingency table.

2. The usage of the sixty-one codons of the I genes can be viewed simply as a two-dimensional contingency table with I rows and sixty-one columns.

3. Synonymous codon usage can be studied by considering the sixty-one sense codons as comprising twenty categorical variables (amino-acids), each with between one and six categories. The two amino-acids with no synonymous codons can however be removed as their usage contributes no information.

Only the second possibility retains information on amino-acid composition. The relationship between synonymous codon usage and amino-acid composition is not completely known and, at an exploratory stage, a study of codon usage may be more informative. There is an additional reason for viewing the data as a two-way contingency table: there are a number of available data exploration/data reduction techniques that are appropriate to this type of data structure.

A minor point to note is that not all organisms use the universal genetic code. The number of sense codons can be sixty, sixty-one, or sixty-two codons. This point will be discussed in section 2.6.

2.3.2. Introduction to Data Reduction.

Before discussing various techniques designed to display patterns in contingency tables, a brief introduction to some aspects of these techniques is appropriate.

A two-way I x J contingency table contains information on the relationships between the J variables and between the I objects. The objective of Data Reduction methods is to find patterns and relationships and thus provide a relatively simple description of the data. These methods usually rely on graphical presentations of the output of the analysis.

The basic features of data reduction will be introduced with the aid of the following example. The study of nucleotide composition is less complex than codon usage. Table 2.1 is the 5 x 4 contingency table representing the usage of the four nucleotides, (U/T,C,A and G), in the third codon position of five hypothetical genes g1 . . g5. This example is adapted from Greenacre (1984).

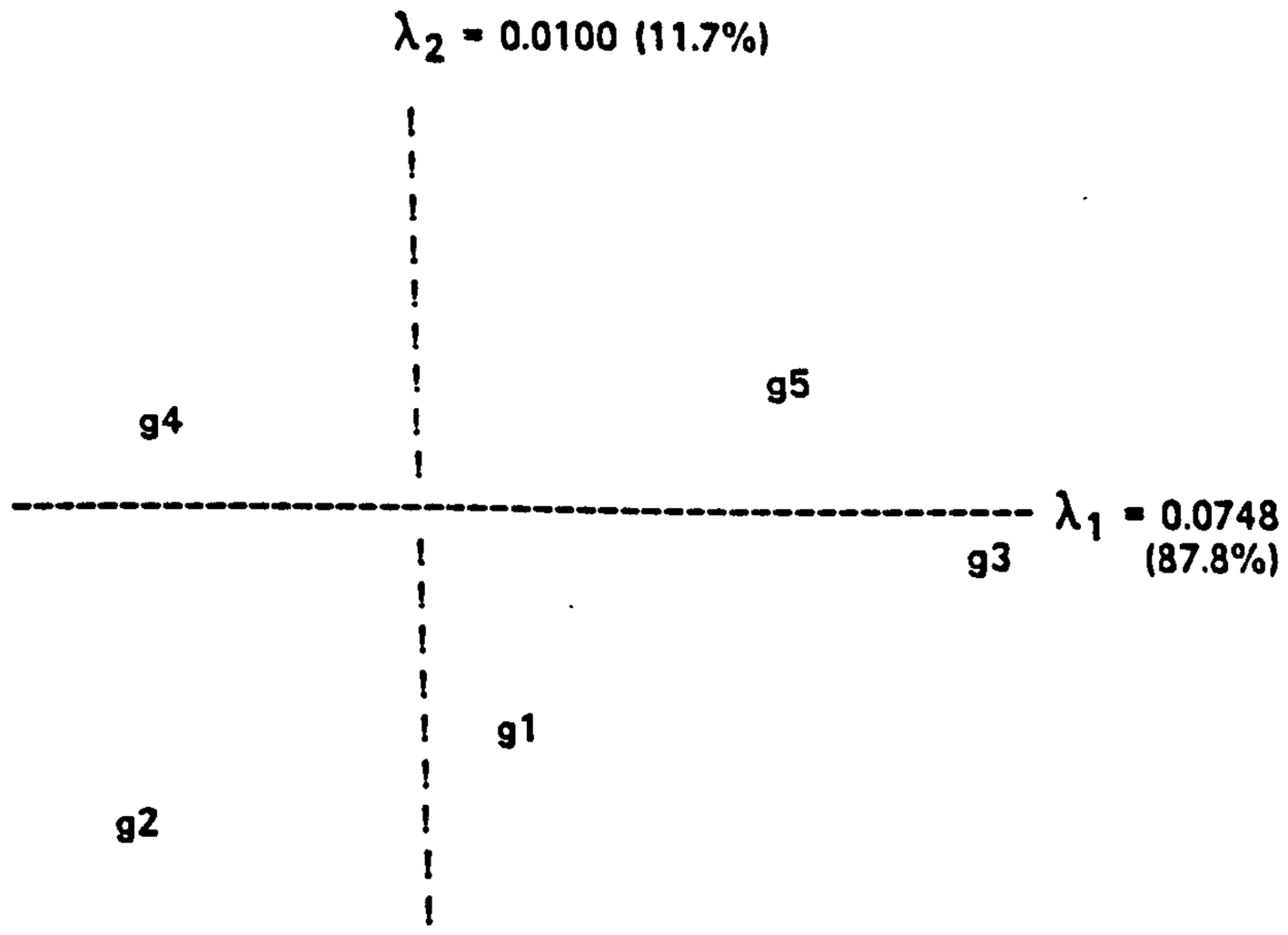
Table 2.1: Nucleotide Usage Example.

		third position nucleotide usage				
		U/T	C	A	G	gene totals
gene	g1	4	2	3	2	11
	g2	4	3	7	4	18
	g3	25	10	12	4	51
	g4	18	24	33	13	88
	g5	10	6	7	2	25
nucleotide totals		61	45	62	25	193

The relative usage of the four nucleotides by each gene, obtained by dividing by the respective gene total, will be called the **profile** of that gene.

If this 5 x 4 data matrix is analysed using an appropriate (see next section) data reduction technique, the gene profiles and nucleotide profiles (i.e. column profiles) will be plotted w.r.t. the new axes. The plot of the gene profiles is shown in figure 2.01:

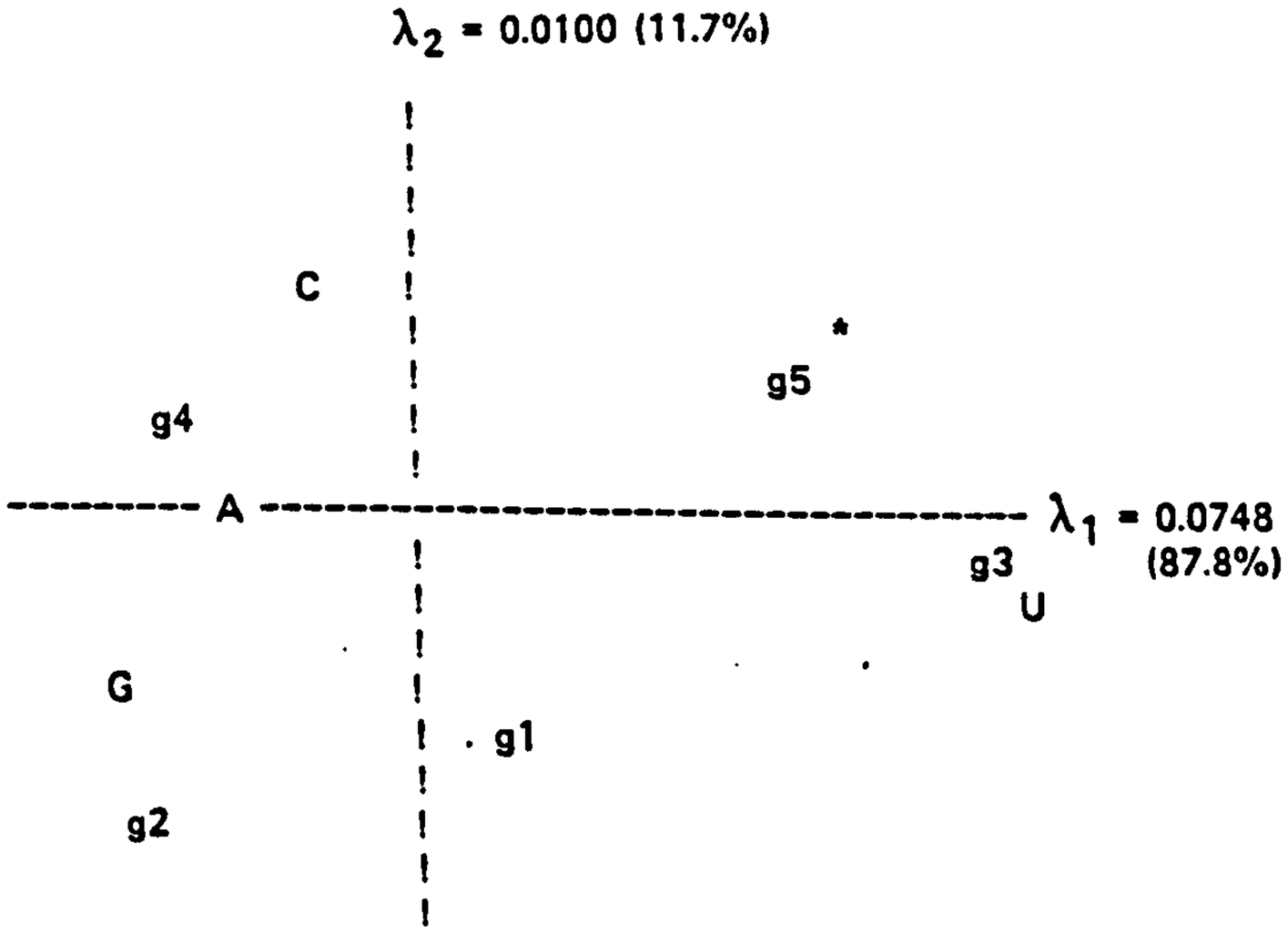
Figure 2.01



The above plot reveals that two of the new dimensions "explain" 99.5% of the spatial variation (i.e. 87.8 + 11.7). The plane defined by these two axes is

thus displaying nearly all the information in the original data matrix. It is clear that genes g3 and g5 are similar in their nucleotide usage, and very different from g2 and g4. The values λ_1 and λ_2 are the eigenvalues of the respective axes. The corresponding nucleotide profiles can be plotted in the same space (see figure 2.02):

Figure 2.02



An initial inspection of this plot allows a description of the axes in terms of nucleotide usage. The first principal axis, which explains 87.8% of the spatial variation, appears to represent the usage of nucleotide U. The second principal axis (11.7%), appears to represent the relative usage of C and G. (The asterisk near gene g5 will be explained in section 2.4.1.)

The output of the data reduction of the original contingency table is the co-ordinates of the gene profiles and nucleotide profiles with respect to the new axes. For each of these new axes, the proportion of the total spatial variation explained can be quantified.

A 5 x 4 contingency table will result in three principal axes being produced (i.e. one less than the smaller dimension of the table). In the above example the third principal axis explained only 0.5% of the total variation. For tables with larger dimensions, such an excellent 2D representation is unlikely. However, it may still be possible to display the main features of the

contingency table in a few dimensions. This will involve a trade-off between ease of interpretation and completeness of description. In a successful application of a data reduction technique the insights gained will far outweigh the loss of information.

2.3.3. Data Reduction Techniques for Contingency Tables.

A family of related methods exists for quantifying patterns in qualitative data: these have been termed "Dual Scaling" techniques by Nishisato (1980). Examples include:

- Reciprocal Averaging
- Guttman Weighting
- Simultaneous Linear Regression
- Canonical Variates Analysis
- Principal Co-ordinates Analysis
- Optimal Scaling
- Biplot
- Principal Components Analysis for Qualitative Data
- Correspondence Analysis

The relationships between some of the various Dual Scaling Methods have been investigated by various authors (Joliffe 1986, Greenacre 1984, Gittins 1985, Nishisato 1980).

Correspondence analysis is ideally suited to the analysis of two-way contingency tables for two reasons. Firstly, no transformation of the data is required (Nishisato 1980). The choice of the data reduction technique depends on the particular interest of the analyst. The I x J contingency table holds information about the I objects (genes) and the J variables (codons). This classification of the two categorical variables as "objects" and "variables", while not appropriate for contingency tables, is commonly used in discussing the analysis of two-dimensional data matrices. For example, analyses can be

categorised as to whether they quantify between-variable variation (e.g. Principal Component Analysis), or between-object variation (e.g. Principal Co-ordinates Analysis), or both. Correspondence analysis quantifies both relationships between objects, and relationships between variables.

All these methods utilise a distance metric in multidimensional space. These distances can be used as an input to a cluster analysis. Grantham's group use cluster analysis (automatic classification) to detect groupings of genes and of codons. Grantham's correspondence analysis/cluster analysis approach is not used here to avoid undue emphasis on discrete as opposed to continuous patterns in the data.

2.3.4. Review of Multivariate Studies of Codon Usage.

As the amount of available sequence data has increased, three correspondence analyses of codon usage patterns from a wide range of genomes and genes have been published (Grantham 1980a,b;1981). The number of genes studied was 90, 119, and 161, respectively. Rowe *et al.* (1984) have carried out a cluster analysis of 332 genes using nucleotide frequency data for the three codon positions. Another correspondence analysis by Grantham's group (Grantham *et al.* 1985) (restricted to host and viral codon usage only) will not be discussed here.

Two papers were initially published reporting a correspondence analysis combined with a Cluster Analysis ("Automatic Classification") based on initially 90 and then 119 genes of different types. Only sequences of over 150 bp were studied. The 61 codons of the "universal code" were used, even though the sample contained some mitochondrial sequences. Differences in amino acid usage were not taken into account, although a separate analysis on the amino acid usage did not reveal a pronounced pattern.

Grantham's group use correspondence analysis in a slightly different way from that used in this chapter. The 61 codons and the genes are plotted in (usually) only two dimensions, and groupings found by cluster analysis are circled. The proportion of the total spatial variation displayed on the plots is not quoted.

The codon points and gene points were plotted for the first two axes, although no value for the proportion of the information displayed was given. The first axis was interpreted as third position G+C content, the second as the relative amounts of U and A in the third position. Cluster Analysis confirmed that certain groups were distinct; these groups being different genomes. Thus each genome (e.g. *E.coli* genes, Mammalian nuclear, etc) contained genes with similar codon usage patterns: most variation between genes was between genome-type not within genome-type. This was Grantham's "Genome Hypothesis".

The third paper (Grantham *et al.* 1981) reported that a further correspondence analysis had been carried out on the 161 sequences then available. The authors noted that it had the "same general structure" as the analysis of 119 genes, but the plot was not presented because of "too many genes for clarity". Discrete within-genome variation was however evident in this analysis (see groups listed below). They also carried out separate analyses on subsets of the available data, and on means of the 14 taxonomic groups involved. These 14 groups were:

ssRNA phages (4 genes).
ssRNA eukaryotic viruses (7 genes).

ssDNA phages (38 genes).

dsDNA phages (13 genes).
dsDNA eukaryotic virus : Papova (12 genes).
dsDNA eukaryotic virus : Hepatitus and Adenovirus (5 genes).

Bacterial (mainly *E.coli*) : Highly expressed genes (13 genes).
Bacterial (mainly *E.coli*) : Weakly expressed genes (16 genes).

Mitochondrial : Yeast (3 genes).

Nuclear : Yeast and Slime Mould (10 genes).

Nuclear/Animal : non-mammal (14 genes).
Nuclear/Animal : mammal excl. human and Ig (11 genes).
Nuclear/Animal : mammal Ig (8 genes).
Nuclear/Animal : mammal human (7 genes).

This analysis of the 14 group means was repeated here to discover how

much of the overall variation was displayed on a plot of the first two principal axes. The first two principal axes displayed 38.7 and 17.4% respectively. (Comparison of the 2D plot produced with Grantham *et al.*'s (1981) Figure 4 served as an additional check on the GENSTAT program used in this chapter). These two axes derived from group total appeared similar to those produced by non-pooled data on earlier analysis by Grantham's group. The first principal axis showed scatter consistent with variation in third position G+C content; the second axis separated bacterial highly expressed genes and yeast & slime Mould genes from the other 12 groups.

The analysis of the 14 group totals emphasised inter-group differences. The other analysis of subsets of the data produced the following results:

1. The 29 Bacterial genes. These split into 3 groups : Highly expressed genes were nearly all contained in one group. This accounted for most of the variation on the first principal axis. The two weakly expressed groups were separated on the second principal axis. Similar results were obtained when the usage of the anticodons was studied (by pooling codon usage data). The second principal axis appears to discriminate according to the use of three amino-acids (Phe & Cys versus Arg).

2. The 79 Bacteriophage and Eukaryotic Viral genes. Three clusters are found. One of these contains mainly ssDNA phage genes.

3. The 84 Bacteriophage and Bacterial genes. Four groups were produced with two Bacteriophage T7 genes (gene 1, and gene 1.1) in the same group as the highly expressed bacterial genes.

4. The 76 Eukaryotic Viral and Eukaryote genes. Four groups produced. The main axis was third position G+C content.

5. The 54 Eukaryotic genes (excluding viruses). Five groups produced. The first principal axis was again G+C content in third base of codon.

Rowe *et al.* (1984) carried out an analysis of 332 genes. Instead of analysing the usage of the 61 codons, the authors chose to concentrate on the nucleotide composition of the three codon positions. This allowed each gene to be visualised in a nine-dimensional "codon space", since each codon

position contributed three degrees of freedom. A cluster analysis then found distinct clusters for (i) vertebrate mitochondria, (ii) invertebrate (yeast) mitochondria, and (iii) phages. A general trend towards high G content, at the expense of A+T, was found in "advanced organisms". Their graphical displays feature only the positions of the cluster centres, and thus even less of the information in the codon usage of the original 332 genes is displayed.

Rowe *et al.* (1984) have criticised the usefulness of the correspondence analysis/cluster analysis approach of Grantham's group in that the bias in codon usage is not related directly to the properties of the sequences. However, their own approach averages out an unquantified amount of the codon usage information and is thus only an analysis of nucleotide frequencies.

The use of correspondence analysis to display the main features of codon usage patterns adds much to our understanding. However, the approach of Grantham's group does not quantify how much of the original information is displayed on the plots produced. Principal axes other than the first two are seldom investigated, even though the third axis may be very similar to the second in the amount of information it displays. The plots produced by such techniques as correspondence analysis still require considerable effort in interpretation: in particular the richness, in information terms, of the plots produced by the third analysis by Grantham's group were never fully highlighted. Finally, the use of cluster analysis may well result in continuous patterns in the data being missed.

2.4. Theoretical Aspects of Correspondence Analysis.

2.4.1. Outline of the Method.

Correspondence analysis is a technique for displaying the rows and columns of a two-way contingency table as points in corresponding low-dimensional vector spaces. Unlike other dual scaling techniques, considerable attention is given to the geometrical aspects of the method. Greenacre & Vrba (1984) describe the main features of correspondence analysis in terms of finding "lines and planes of closest fit" to a cloud of

points in multidimensional space, based on Pearson's original approach to Principal Components Analysis (Pearson 1901). The approach used here is based on Greenacre & Vrba (1984) and Greenacre (1984). Correspondence analysis has five main features:

The cloud of points in multidimensional space.

The codon usage of I genes (using the universal genetic code) can be represented as an I x 61 contingency table. This data matrix, denoted **N**, can be viewed as I rows representing the I genes. Each row can be corrected for gene length by dividing by the respective row total, thus yielding the gene's co-ordinates in the 61-dimensional codon space defined by the sixty-one sense codons. Such a matrix containing the co-ordinates of the rows (genes) in codon space is denoted **R**.

The position of a gene's codon usage profile in the 61-dimensional space is therefore independent of the total length of the gene. However the influence of a particular gene on the positioning of the principal axes by correspondence analysis is directly proportional to its length. The actual quality, the **mass**, is not the length, but the length divided by the total number of codons in the contingency table.

The centre of the cloud of I genes in 61-dimensional codon space is simply the average codon usage profile of the I genes i.e. the total codon usage of all I genes divided by I. This is termed the **centroid**, or centre of mass of the cloud.

The above three concepts –the position of a gene, the mass of a point, and the centroid – can be phrased succinctly in matrix notation. The original contingency table **N** is first converted into a matrix of relative frequencies, **P**, by dividing each element of **N** by the grand total of the elements of **N**.

The row and column totals of **P** are denoted, respectively, by:

$$r_i = \sum_{j=1}^{61} p_{ij} \qquad i = 1,..,I \tag{2.1}$$

$$c_j = \sum_{i=1}^{I} p_{ij} \qquad j = 1,..,61 \tag{2.2}$$

These row and column totals can be represented by the vectors **r** and **c** respectively:

$$\mathbf{r} = [\, r_1 \; r_2 \; r_3 \; \cdots \; r_I \,] \tag{2.3}$$

$$\mathbf{c} = [\, c_1 \; c_2 \; c_3 \; \cdots \; c_{61} \,] \tag{2.4}$$

These row and column totals are usually held in two diagonal matrices $\mathbf{D_r}$ and $\mathbf{D_c}$ respectively.

$$\mathbf{D_r} = \text{diag} \{ \mathbf{r} \} \tag{2.5}$$

$$\mathbf{D_c} = \text{diag} \{ \mathbf{c} \} \tag{2.6}$$

The positions of the I genes in 61-dimensional codon space are therefore contained in rows of the I x 61 matrix **R**:

$$\mathbf{R} = \mathbf{D_r^{-1}} \, \mathbf{P} \tag{2.7}$$

The masses of these I genes are given simply by the elements of **r**, and the position of the centroid is given by the elements of **c**.

The metric structure of space.

Given a set of I genes, it is now possible to describe the position of their codon usage profiles in 61-dimensional space, and to assign masses to them according to the relative gene lengths. The centroid of this cloud of genes is simply the weighted average of the I codon usage profiles. The distance between two genes (or codon usage profiles) is now required. A simple Euclidean distance would give undue emphasis to codons that were relatively

highly used in all genes. Instead correspondence analysis corrects for this by dividing the individual parts of the squared difference term by the elements of the average codon usage profile, **c**. Denote the point in 61-dimensional codon space representing the i^{th} row of **R** (i.e. the i^{th} gene) by:

$$a_i = [\ p_{i1}/r_i \quad p_{i2}/r_i \quad \cdots \quad p_{i61}/r_i\ ] \tag{2.8}$$

The distance between two genes $\mathbf{a}_1$ and $\mathbf{a}_2$ is therefore:

$$d^2(\mathbf{a}_1,\mathbf{a}_2) = (\mathbf{a}_1 - \mathbf{a}_2)^T\ \mathbf{D}_\mathbf{c}^{-1}\ (\mathbf{a}_1 - \mathbf{a}_2) \tag{2.9}$$

$$= \sum_{j=1}^{61} \left((p_{1j}/r_1) - (p_{2j}/r_2) \right)^2 / c_j \tag{2.10}$$

where $\mathbf{D}_\mathbf{c}$ is a 61 x 61 diagonal matrix of the c_j's (as defined above). This distance is proportional to a χ^2 statistic, and is commonly known as the "χ^2 distance". The proportionality factor in the above example is the combined length of the two genes.

The goodness-of-fit of this space to a subspace.

The purpose of correspondence analysis is to produce a set of principal axes that will allow a low-dimensional display to be produced that will still contain the main features of the data. The production of these principal axes requires a measure of "closeness" of a line to the cloud of points in multidimensional space. The measure of goodness-of-fit used is the weighted sum of squared distances from the points to the line. This involves the use of the row masses (as weights) and the χ^2 distance as detailed above. The generation of the principal axes can be thought of as an iterative process by which the best fitting line is chosen as the first new dimension. The process is then repeated to find the second and subsequent axes for the remaining "variation" in the cloud.

The details of the matrix algebra underlying the generation of the principal axes can be found in Greenacre (1984). Greenacre & Vrba (1984) note

that the principal axes can be obtained by solving the eigen-equation:

$$(P^T D_r^{-1} P D_c^{-1})u = \lambda u \qquad (2.11)$$

where:

D_r and D_c are the diagonal matrices of the row and column masses respectively, and u denotes the eigenvectors, and λ the eigenvalues.

A computationally easier method of obtaining the principal axes, the **singular value decomposition** (SVD), will be outlined in section 2.4.3. It is noted here however that the use of the SVD method produces the co-ordinates of each gene's codon usage profile w.r.t. the principal axes. This I x 60 matrix is denoted F, the equivalent 61 x 60 matrix for the codon profiles is denoted G.

The first eigenvalue represents a trivial solution in which the first principal axis is the centroid. This effectively subtracts one dimension from the space. The remaining eigenvalues are a measure of the proportion of the "variation" explained by the principal axes. This variation can be quantified: it is simply the weighted sum of squared distances of points from the centroid. This **total inertia** is however identical to the sum of the non-trivial eigenvalues (which are also referred to as **principal inertias**).

The symmetry of the rows and columns.

In section 2.3.3 it was noted that correspondence analysis allowed the study of patterns between "objects" and also between "variables": i.e. both categorical variables are analysed. So far only the analysis of the row profiles (codon usage profiles) has been discussed. Note that the original data matrix (contingency table) could easily be transposed and the analysis of the row details detailed above applied to column profiles. However, both analysis of row profiles and column profiles are carried out simultaneously in a correspondence analysis. There exist two clouds of points: one the row profiles (codon usage profiles); the other the column profiles (codon profiles). The relationship between these two spaces is intimate: they have the same

set of eigenvalues and hence the same total inertia. This means that points representing the I genes and the 61 codons can be put on the same plot. If a gene point and a codon point are found to lie closer than would be expected under a simple 2D random model, it can be inferred that the gene in question has high usage of that codon.

The two matrices of co-ordinates produced by a correspondence analysis are the F matrix (row (gene) co-ordinates w.r.t. the principal axes), and the G matrix (column co-ordinates w.r.t. the principal axes).

Transition formulae and supplementary profiles.

The i genes with their usage of the 61 sense codons determine directly the 60 new non-trivial principal axes. However, hypothetical points can be plotted w.r.t. principal axes. For example, a hypothetical gene with all 61 codons used equally can be plotted along with the data used to derive the principal axes. These additional points (or profiles) are termed **supplementary profiles**, and can be considered as similar to other points except that they have no mass. The co-ordinates w.r.t. the principal axes are calculated using **two transition formulae**. The derivation of the formulae are not discussed here but they arise from the relationship between the two clouds of points (see Greenacre 1984). The codon usage of the hypothetical gene can be considered as a row vector **q** where the sum of the usage is unity. The co-ordinate of **q** w.r.t. the k^{th} principal axis requires information on the k^{th} column of **G**, the 61 x 61 matrix of co-ordinates of the 61 codons in the new space, and the value of the k^{th} eigenvector, λ_k. The required co-ordinate is then:

$$q\, g_k\, \lambda_k^{-1/2} \qquad (2.12)$$

where g_k is the k^{th} column of G.

An analogous transition formula facilitates the plotting of a column profile (e.g. the relative usage of a particular codon in the set of genes studied), but is of little use in this study.

2.4.2. Application to the Nucleotide Usage Example.

Some of the above features of correspondence analysis can be illustrated using the nucleotide usage example outlined in section 2.3.2. The 5 x 4 contingency table (see table 2.1) was analysed using correspondence analysis and the co-ordinates of the gene profiles and codon profiles displayed in figures 2.01 and 2.02. The output of this analysis is shown in table 2.2.

Table 2.2

Co-ordinates

		1	2	3	
	g1	0.066	-0.194	0.071	
	g2	-0.259	-0.243	-0.034	
Gene	g3	0.381	-0.011	-0.005	Matrix F
Profile	g4	-0.233	0.058	0.003	
	g5	0.201	0.079	-0.008	
	U/T	0.393	-0.030	-0.001	
Nucleotide	C	-0.099	0.141	0.022	Matrix G
Profile	A	-0.196	0.007	-0.026	
	G	-0.294	-0.198	0.026	

	1	2	3
Eigenvalues:	0.07476	0.01002	0.00041

Note that the number of principal Axes is one less than the smallest dimension of the original contingency table due to centering the cloud. The total inertia of both clouds is simply the sum of the non-trivial eigenvalues (i.e. 0.08519). The first two principal axes explain nearly all of this spatial variation (0.08478/0.08519 = 99.5%). The first two principal axes were interpreted in section 2.3.2 as:

Principal Axis 1 - Use of U Nucleotide.

Principal Axis 2 - Relative Use of C and G Nucleotides.

The number of genes, five, represents a very small sample. Assume that the average nucleotide usage of a much larger sample was known (e.g. the percentage usage of U/T:C:A:G = 42 : 29 : 20 : 9), then it would be interesting to see where this point lay on the 2D display shown in figure 2.02. To plot this point as a supplementary profile equation (2.12) is used. This is simply the product of a row vector representing the point to be plotted and the relevant column of the matrix **G** (see Table 2.2), divided by the square root of the relevant eigenvalue. To calculate the co-ordinate on the first principal axis, the values of **q**, **g** and λ used are:

$$\mathbf{q} = [\ 0.42 \quad 0.29 \quad 0.20 \quad 0.09\ ]$$

$$\mathbf{g}^T = [\ 0.393 \quad -0.099 \quad -0.196 \quad -0.294\ ]$$

$$\lambda = 0.07476$$

Inserting these values into equation (2.12) yields a value of 0.258. The co-ordinate of this supplementary profile on the second principal axis is simply the matrix product of **q** with the second column of the matrix **G**, divided by the square root of the second eigenvalue. The co-ordinates of the supplementary point w.r.t. the first two principal axes is therefore [0.258 0.118]. This point is marked with an asterisk in Figure 2.02 and lies close to gene 5. This suggests that gene 5 has a nucleotide usage pattern similar to that of the average of the larger sample.

2.4.3. The Singular Value Decomposition.

A computationally easy method of obtaining the principal axes is available with most modern statistical packages, and high-level statistical programming languages (e.g. GENSTAT; see section 2.5.3). The **singular value decomposition (SVD)** is a general matrix algebra technique that includes the eigenvalue/eigenvector decomposition as a special case. Its use in Correspondence analysis is lucidly explained in Greenacre (1984). In this section, only the matrix computations that produce the co-ordinates of the gene and codon profiles (matrices F and G respectively) w.r.t. the principal axes are outlined.

Before defining the SVD of a real I x J matrix **A**, some common definitions of terms used in matrix algebra are appropriate:

- The **rank** k of an I x J matrix **A** is a measure of the dimension of **A**. If some of the rows (or columns) are not independent, then k will be less than I and J, whichever is smaller. If this were the case for the correspondence matrix **P**, then the total inertia could be displayed in a space with a dimension less than 60.

- The codon usage profile of a gene, as noted earlier, is described in terms of its co-ordinates w.r.t. 61 dimensions specified by the sense codons. The 61 sense codons thus represent vectors in 61 dimensional codon space. The position of any gene is given in terms of co-ordinates w.r.t. these **basis vectors**. Consider the cloud of genes in this multidimensional space as fixed. If the axes are rotated, a new **basis** is obtained where the new **basis vectors** are linear combinations of the original **vectors**.

The SVD can now be outlined in a non-rigorous fashion. For a comprehensive treatment, consult Golub & van Loan (1983). The SVD of a real matrix **A** of rank k is an expression of the form:

$$\underset{i\times j}{\mathbf{A}} = \underset{i\times k}{\mathbf{U}} \quad \underset{k\times k}{\mathbf{D}_{\alpha}} \quad \underset{k\times j}{\mathbf{V}^{T}} \qquad (2.13)$$

where D_α = diag { α_1 α_2 . . . α_k } is a diagonal matrix (see below) and **U** and **V** contain no dependent columns, or rows, respectively.

The k column vectors of **U** (the left singular vectors) form a basis for the columns of **A**. They are also the eigenvectors of AA^T, with associated eigenvalues equal to the squares of the elements of D_α (the singular values). Note the similarity with equation (2.11). Similarly the k column vectors of **V** (the right singular vectors) form a basis for the rows of **A**, and are also the eigenvectors of A^TA. The associated eigenvalues are again the squared singular values.

The ordinary SVD method cannot immediately be applied to the correspondence matrix **P** due to the particular weighting of the profiles and the distance metric that are used in correspondence analysis. A solution to

this problem is to suitably weight P to produce a matrix A. After obtaining the SVD of A, the U and V matrices can be "corrected" to give the appropriate matrices for P (these are usually labelled N and M respectively). N (I x K), and M (J x K) define the principal axes w.r.t. the rows, and columns respectively, of the original contingency table. From N and M the co-ordinate matrices F (I x K) and G (J x K) are obtained. The singular values of P are equal to those of A.

The actual steps are as follows:

(1) Let $A = D_r^{-1/2} P D_c^{-1/2}$ (2.14)

(2) Find the ordinary SVD of A: $A = U D_\alpha V^T$ (2.15)

(3) Let $N = D_r^{-1/2} U$, $M = D_c^{-1/2} V$ (2.16)

The eigenvalues associated with the correspondence matrix P are obtained by squaring the singular values that are contained in D_α.

(4) The co-ordinates of the gene profiles and the codon profiles are held in F and G respectively:

$$F = ND_\alpha, \quad G = MD_\alpha \tag{2.17}$$

These steps were implemented as a GENSTAT program (see section 2.5.3). Thus for a given contingency table N, the co-ordinate matrices F and G plus the non-trivial eigenvalues were produced.

2.5. Data Sources, Data Manipulation and Data Analysis.

2.5.1. Nucleotide Sequence Data Libraries.

Nearly all the DNA sequence data used in the correspondence analysis was extracted from the EMBL Nucleotide Sequence Data Library (release 2; April 1983), which was obtained from EMBL in Heidelberg (FRG) on magnetic tape and loaded on to the ICL 2972 mainframe at Edinburgh University. The NIH GenBank data library was used to a lesser extent. The GenBank data library and later releases of the EMBL data library became available on a VAX 8500 mini-computer (Edinburgh University ERCVAX) during the later stages of

this research. Some sequence data was obtained directly from the literature: these are noted in section 2.6).

The rate of increase in size of nucleotide sequence data libraries has been, and continues to be, impressive. A comparison of releases 2 (April 83) and release 10 (December 86) of the EMBL data library is made in Table 2.3.

Table 2.3

EMBL Data Library Growth

Sequences	Release 2 (Apr 83) Entries	Bases	Release 10 (Dec 86) Entries	Bases
Artificial	5	8238	192	71237
Chloroplast	12	12616	167	465378
Genetic elements	18	18341	54	43857
Mitochondrial	39	109008	322	376392
Prokaryotic	103	134270	1175	1305116
Viral/Phage	171	391292	1335	1975030
Eukaryotic	463	440682	5540	5465260
Unclassified	0	0	32	64678
Total	811	1114447	8817	9766948

Not all the entries in the EMBL and GenBank data libraries contain sequence data that codes for amino-acids. At the start of this research, only the EMBL data library quoted the known start and stop positions of protein-coding DNA/RNA within a given entry. Each EMBL data library entry contained formatted additional information prior to the listing of the actual sequence data. Thus computer programs could automatically pick up the location of the coding portions before reading in the sequence. This allows considerable automation in the production of codon usage tables.

For these reasons, the initial selection of protein-coding sequence data was made from the EMBL data library. It should be noted that the actual entries do not constitute an unbiased sample of all coding DNA, but reflect research interests spread over many areas of biology.

2.5.2. Sequence Manipulation Software.

Before any statistical analysis of sequence data could be undertaken, there was a requirement for a set of computer programs to check the data for errors, to produce codon usage tables and related summaries, and to allow general manipulation of sequence data. Software which runs on this processed data is outlined in section 2.5.3 (Data Analysis/Statistical Software).

Initially a set of FORTRAN77 programs and subroutines were written to process the entries in the EMBL data library. The existence of errors in the stated start/stop positions of coding regions and of deletions/insertions in the actual sequence data imposed a requirement for a series of checks before processing could continue. These checks included:

- a test that the length (in amino-acids) of the coding region was a positive integer. Failure resulted in a thorough manual check of the additional information provided with the EMBL data entry and, if necessary/possible, the original reference.

- a test that the codon usage table contained no more than one stop codon. As above, manual checks were then carried out.

The software read in the standard EMBL format and could produce singlet, doublet and triplet tables for all three possible reading frames. Given the start/stop points for a coding region, it could also provide this summary information for flanking regions and introns. Although only the codon usage data is used in this study, these programs have been used for other analyses (Weir 1985, Lathe 1985). Only a brief outline of the software has been given here as sequence manipulation software packages are now in common use.

In the later stages of this study, the purchase by the Genetics and Molecular Biology Departments of a comprehensive sequence analysis package (The UWGCG package or 'Wisconsin' package - Devereux *et al.* 1984) provided another set of software. This package, like the later versions of the EMBL and GenBank data libraries, was loaded on to the ERCVAX. The concentration of sequence software and information on the VAX lead to the redundancy of much of the ICL 2972-based FORTRAN77 sequence manipulation software.

2.5.3. Data Analysis/Statistical Software.

A simple GENSTAT correspondence analysis program is listed in appendix B of Greenacre (1984). This was used as a basis for the analysis used here. GENSTAT is a high-level statistical language that greatly simplifies the programming of matrix computations. For example, an ordinary SVD can be computed in a one line statement. The input to the program was the contingency table of 428 genes x 61 codons. The program logic followed the steps outlined in section 2.4.3. The main output was the co-ordinates of both the genes (a 428 x 60 matrix), and the codons (a 61 x 60 matrix), w.r.t. the principal axes plus the 60 non-trivial eigenvalues.

Part of the output of the GENSTAT program was used as input to a small FORTRAN77 program that produced supplementary profiles as required. All the graphical displays were produced using EASYGRAPH, a plotting program developed at Edinburgh University.

The transition formula used in the supplementary profile FORTRAN77 program allows sequences that did not take part in the analysis to be plotted w.r.t. the principal axes. This facility can be used as a check on the accuracy of the output from a correspondence analysis.

The co-ordinates of the 428 genes were re-calculated from their original codon usage patterns by use of the eigenvalues and the co-ordinates of the 61 codons w.r.t. the principal axes (i.e. the G matrix). The GENSTAT program was set up to write out the eigenvalues and the G matrix to five and three decimal places respectively. Comparisons between the "reconstituted" gene co-ordinates and those output by the analysis showed that there was good agreement (to within one per cent).

2.6. Sequence Data Used in the Analysis.

Most of the data was obtained from the GenBank and EMBL databases, although gene types of special interest were typed in from recent references (e.g. *B.subtilis*, *N.crassa*). Genes less than 150 bp were excluded, as were genes of dubious coding ability. Species were usually chosen to represent larger taxonomic classes e.g Human, for mammals. Although only one large

analysis was carried out, the results are plotted over five graphs to aid interpretation. These are:

1	Eukaryotic Nuclear (EN) Genes	9 subgroups (108 genes).
2	Eukaryotic Viral (EV) Genes	9 subgroups (51 genes).
3	Eukaryotic Organelle (EO) Genes	7 subgroups (44 genes).
4	Prokaryotic (P) Genes	7 subgroups (104 genes).
5	Bacteriophage (B) Genes	7 subgroups (121 genes).

Further details are given in Table 2.4. A full list of the 428 genes in the study is contained in the Appendix.

The choice of sequence data for an initial study of codon usage patterns across all species was guided be the following:

1. The analysis had an upper limit of 429 sequences due to the 256k workspace required by the GENSTAT package as implemented on the ICL 2972 mainframe at Edinburgh University.

2. An attempt was to be made to reduce the bias in the available data towards certain species.

3. Single sequences for one species were excluded unless they could be lumped with a similar genome type.

4. Considerable care was to taken to include sequences definitely known to be coding. Sequence data containing errors was checked, when reasonably possible, against the original reference.

The analysis was carried out on the 61 sense codons according to the "universal" genetic code. The 24 mammalian mitochondrial genes, the 5 yeast mitochondrial genes and the 4 *Drosophila melanogaster* mitochondrial genes use genetic codes slightly different from the universal genetic code. There is also evidence that ciliated protozoa and a bacterium (*Mycoplasma capricolum*) also use slightly different codes (Fox 1985). Six of the eight protozoan genes in the analysis are from ciliated species (*Tetrahymena thermophila, Stylonychia*

lemnae, *Paramecium tetraurelia*, *Paramecium primaurelia*). Four *Mycoplasma capricolum* genes were studied.

All three of the mitochondrial genomes studied use a universal genetic code Ile codon (AUA) as a second Met codon, and a universal terminator codon (UGA) as a second Trp codon. UGA is also used by *M.capricolum* as a Trp codon. Mammalian mitochondria also use the two universal Arg codons (AGA, AGG) as stop codons; Yeast mitochondria use four universal Leu codons (CUN) as Thr codons; and Drosophila melanogaster also uses two universal Arg codons (AGA, AGG) as Ser codons (see Jukes 1983). Ciliated protozoan species appear to use UAA (a universal stop codon) as a Gln codon; UAG also appears to be a Gln codon in *Paramecium* (see Fox 1985).

The decision to analyse these 43 genes as if they conformed to the universal genetic code allows the combined analysis of 428 genes regardless of slight differences in their genetic codes. This allows comparisons between nuclear and organelle codon usage patterns. The effect of genetic code differences on the results of the analysis will be discussed in section 2.7.

Table 2.4: Summary of Sequence Data Analysed.

Eukaryotic Nuclear subgroups (see figures 2.08 and 2.13).

EN_1 - Mammalia (*H.sapiens*; N=35)
EN_2 - Aves (*G.gallus*; N=9).
EN_3 - Pisces (six species; N=10)
EN_4 - Amphibia (*Xenopus laevis*; N=7).
EN_5 - Arthropoda (*Drosophila melanogaster*; N=11).
EN_6 - Echinodermata (*Psammechinus miliaris*; N=5).
EN_7 - Protozoa (six species; N=8).
EN_8 - Fungi (*Saccharomyces cerevisiae*; N=16).
EN_9 - Spermatophyta (five species; N=7).

Eukaryotic Virus subgroups (see figures 2.09 and 2.14).

EV_1 - Large (172kb) dsDNA Animal Virus (Epstein-Barr Virus; N=3).
EV_2 - Medium (36kb) dsDNA Animal Virus (Adenovirus type 2; N=22).
EV_3 - Small (5kb) dsDNA Animal Virus (SV40; N=5).
EV_4 - ssDNA Animal Virus (Mouse Minute Virus; N=2).
EV_5 - ssRNA Animal Virus (Polio virus).
EV_6 - ssRNA Retrovirus (three retroviruses; N=3).
EV_7 - dsRNA Animal Virus (SAII virus; N=2).
EV_8 - dsDNA Plant Virus (Cauliflower Mosaic Virus; N=5).
EV_9 - ssRNA Plant Viruses (three species (TMV, TYMV, AMV); N=3).

Eukaryotic Organelle subgroups (see figures 2.10 and 2.15).

EO_1 - Mammalian mtDNA I (*H.sapiens* mtDNA; N=8).
EO_2 - Fungal mtDNA (*Saccharomyces Cerevisiae* mtDNA; N=5).
EO_3 - Spermatophyte mtDNA (two mtDNA genomes; N=2).
EO_4 - Plant and Algal cpDNA (eight cpDNA genomes; N=9).
EO_5 - Arthropod mtDNA (*D.melanogaster* mtDNA; N=4).
EO_6 - Mammalian mtDNA II (*M.musculus* mtDNA; N=8).
EO_7 - Mammalian mtDNA III (*B.taurus* mtDNA; N=8).

Prokaryotic subgroups (see figures 2.11 and 2.16).

P_1 - Enterobacterium Ia (*E.coli* chromosomal genes; N=50).
P_2 - Enterobacterium Ib (*E.coli* plasmid and transposon genes; N=8).
P_3 - Enterobacterium II (*Salmonella typhimurium* genes; N=11).
P_4 - Rhizopoda (four Rhizopoda species; N=8).
P_5 - Cyanobacteria (*Anabaena* 7120; N=4).
P_6 - Bacilli (*Bacillus subtilis*; N=19).
P_7 - Mycoplasmas (*Mycoplasma capricolum*; N=4).

Bacteriophage subgroups (see figures 2.12 and 2.17).

B_1 - Large (49kb) dsDNA *E.coli* phage I (λ; N=46).
B_2 - Large (40kb) dsDNA *E.coli* phage II (T7; N=41).
B_3 - Large (170kb) dsDNA *E.coli* phage III (T4; N=11).
B_4 - Small (4kb) ssRNA *E.coli* phages IV (MS2, $Q\beta$; N=5).
B_5 - Small (5kb) ssDNA *E.coli* phage V (ϕX174; N=9).
B_6 - Large (19kb) dsDNA *B.subtilis* phage (ϕ29; N=5).
B_7 - Large (40kb) dsDNA *S.typhimurium* phage (P22; N=4).

2.7. Results from Correspondence Analysis.

The interpretation presented here is based mainly on the approach of Greenacre & Vrba (1984), although an aspect of Greenacre's (1984) approach is also used. The codon usage data of 428 genes was structured as a 428 x 61 contingency table. A correspondence analysis of this data matrix produced 60 non-trivial principal axes each with an associated eigenvalue and the co-ordinates of the 61 codons and the 428 genes w.r.t. these principal axes. A small number of these principal axes were then used to display the main features of the data. The main features of the display were obtained through consideration of the positioning of the genes and codons w.r.t. these principal axes. An additional aid to the interpretation of the display was the plotting of extra points to locate the position of known codon usage patterns.

2.7.1. Total Inertia and Graphical Display.

The Proportion of the Total Inertia due to Amino-Acid Usage.

The total inertia of the codon usage profiles is given by the sum of the 60 non-trivial eigenvalues. To quantify the proportion of the total inertia due to amino-acid composition, a second analysis was carried out on the 428 x 20 contingency table of amino-acid usage. This total inertia of this data matrix represents the effect of removing the effect of synonymous codon usage on the total inertia of the main analysis.

```
Total Inertia (codon usage analysis)  = 0.68427

Total Inertia (a.a.  usage analysis)  = 0.18733
```

This suggests that 27.4% of the total inertia in the main analysis is due to amino acid usage.

The Number of Principal Axes Required to Display the Main Features.

The object of the correspondence analysis is to display the main features of the contingency table. This involves a trade-off between ease of interpretation and completeness of the display. There are no criteria available for choosing how many axes to display. The percentage contribution of each of the 60 non-trivial principal axes to the total inertia is shown as a histogram in figure 2.03. The axes are automatically ordered in order of eigenvalue magnitude.

There is an apparent discontinuity between the proportion of inertia explained by axis 4 and axis 5. From axis 5 onwards, there is a gradual fall. The percentage of the total inertia accounted for by axes 4 and 5 is 6.1 and 3.7 respectively. The first 4 axes explain 43.6% of the total inertia, and the plane of axes 1 + 2 accounts for 30.4%. These first four principal axes were used to display the main features of the codon usage data for the 428 genes.

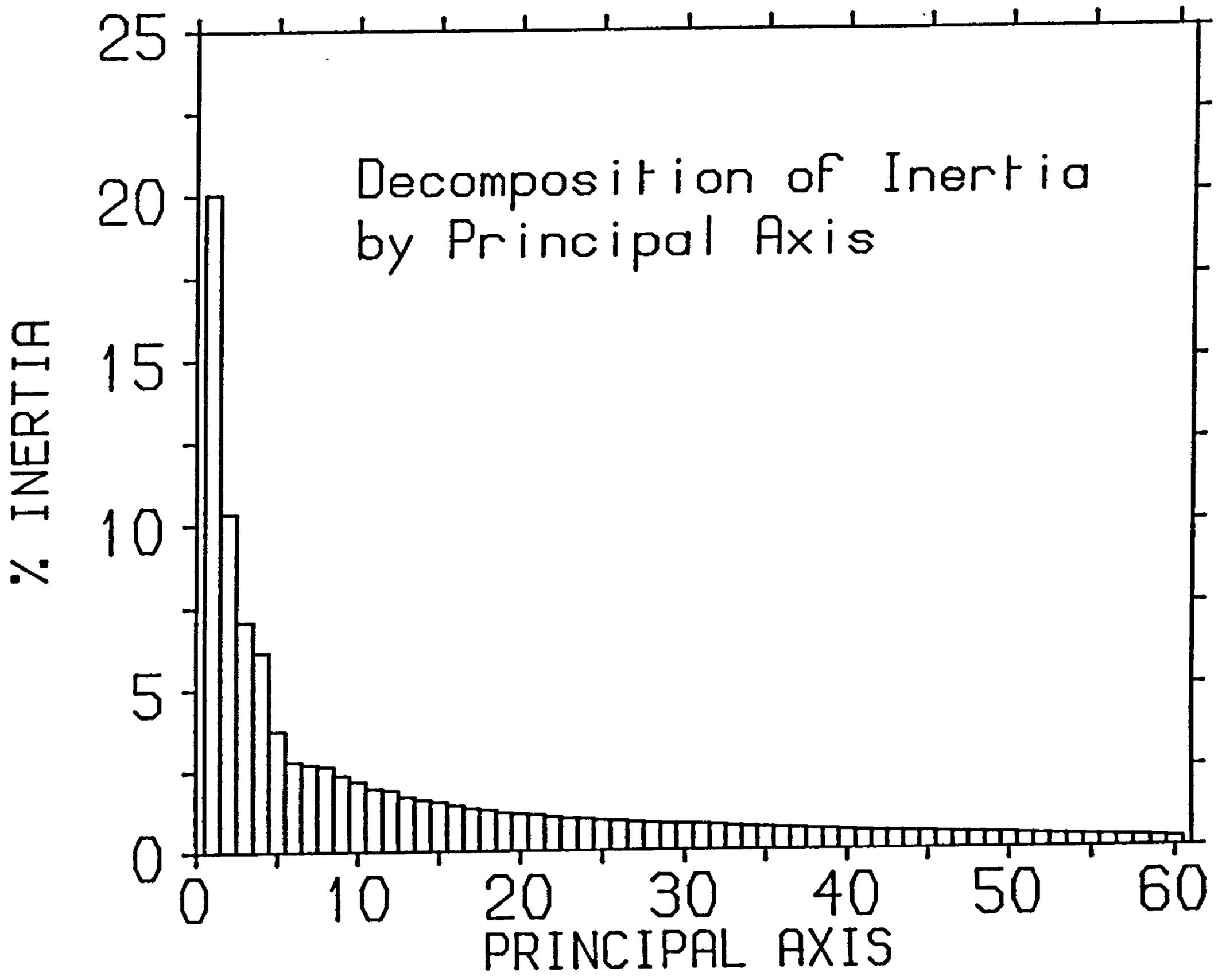
Displaying the Results.

Even a display involving only four principal axes is difficult to present in a digestible form when the number of genes studied is large. To simplify the interpretation, the 428 genes were plotted as five separate groups as outlined in section 2.6. The 61 codons were however plotted on the one graph.

An initial view of the output of the analysis is shown in Figures 2.04 to 2.07. These show all 428 genes plotted together in 2D plots showing the first four principal axes. The 61 codons are plotted similarly. Only very vague ideas of scatter can be obtained from these graphs, as it is difficult to label the respective genes and codons in a clear manner. The genes are replotted, as noted above, over five graphs. Ten graphs are therefore required. The genes are further classified according to the species or taxonomic group to which they belong (see Keys on Figures 2.08 to 2.17). As an aid to interpreting the codon plots (Figures 2.06 and 2.07) the co-ordinates of each of the 61 (universal genetic code) sense codons w.r.t. the first four principal axes are shown in Table 2.5 on page 63. An easier way of interpreting the display is by

Figure 2.03

The percentage contribution of each of the sixty non-trivial eigenvalues to the total inertia (see page 55).


Decomposition of Inertia
by Principal Axis
% INERTIA
25
20
15
10
5
0
0
10
20
30
40
50
60
PRINCIPAL AXIS

plotting additional points on the displays using the transition formulae given by equation (2.12) in section 2.4.1.

2.7.2. Identification of the Principal Axes.

In common with other data reduction techniques, there is no guarantee that the principal axes will be easily interpretable in a biological manner. The dataset studied, while it encompasses codon usage data from a wide range of genomes, is biased. In particular, several species are over-represented: e.g. *E.coli*, bacteriophages T7 and λ; whereas the Plant Kingdom is very under-represented.

The influence that a particular gene (or codon) has had on determining the principal axes is proportional to its squared distance from the centroid, and to the length of the gene (or relative frequency of the codon). The distance from the centroid is therefore a more important consideration than gene length (codon relative frequency), although a cluster of genes (or codons) can obviously exert some influence even if they are comparatively close to the centroid. A consideration of those codons and those genes with high or low values for a given principal axis will therefore aid the interpretation of that axis.

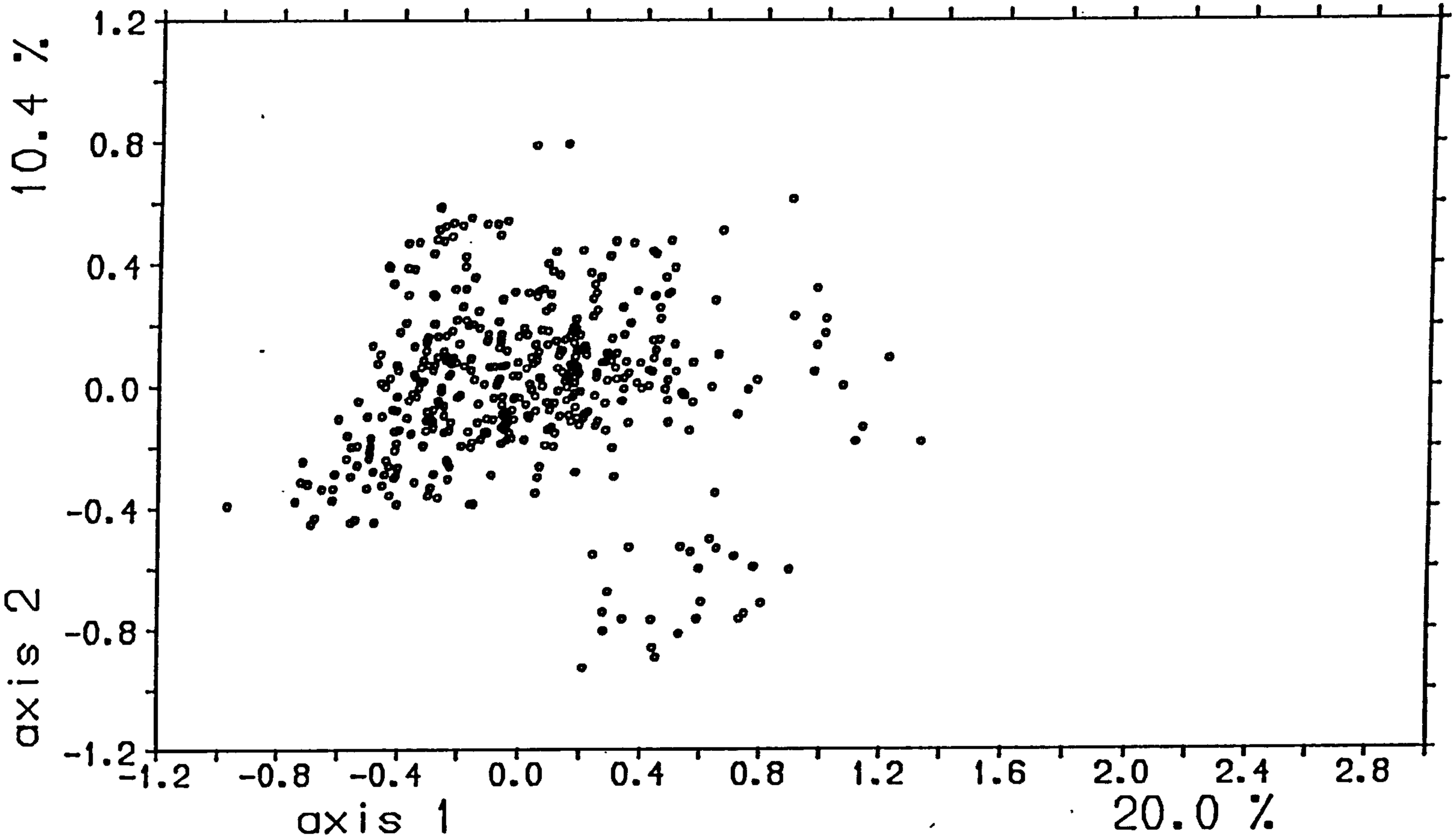
Information from Codon Plots.

Using the codon co-ordinate data from Table 2.5 on page 63 (and Figures 2.06 and 2.07), it is clear that the first principal axes is correlated with G+C content in the third codon position. All but one of the -G or -C ending codons have negative values for this axes, whereas all but one of the -A and -T ending codons have positive values. There is a suggestion that -G ending codons have higher absolute values than -C ending codons. This first principal axis accounts for 20% of the total inertia. The co-ordinate values of the 61 codons for principal axes 2,3, and 4 do not appear initially to be easily classifiable. However the large negative values for codons AGA and AGG on axes 3 and 4 suggest that this axis may be picking out differences between the mammalian mitochondrial genetic code and the other genetic codes. These two codons serve as stop codons in mammalian mitochondria. Amino-acid usage may also be important: both cysteine codons (UGU and UGC) have low values for axes 3 and 4, as does tryptophan (UGG) for axis 4.

Opposite page 59

Figure 2.04

All 428 genes involved in the correspondence analysis plotted according to their co-ordinates on principal axes one and two. Note that this two-dimensional plot displays 30.4% of the total inertia.

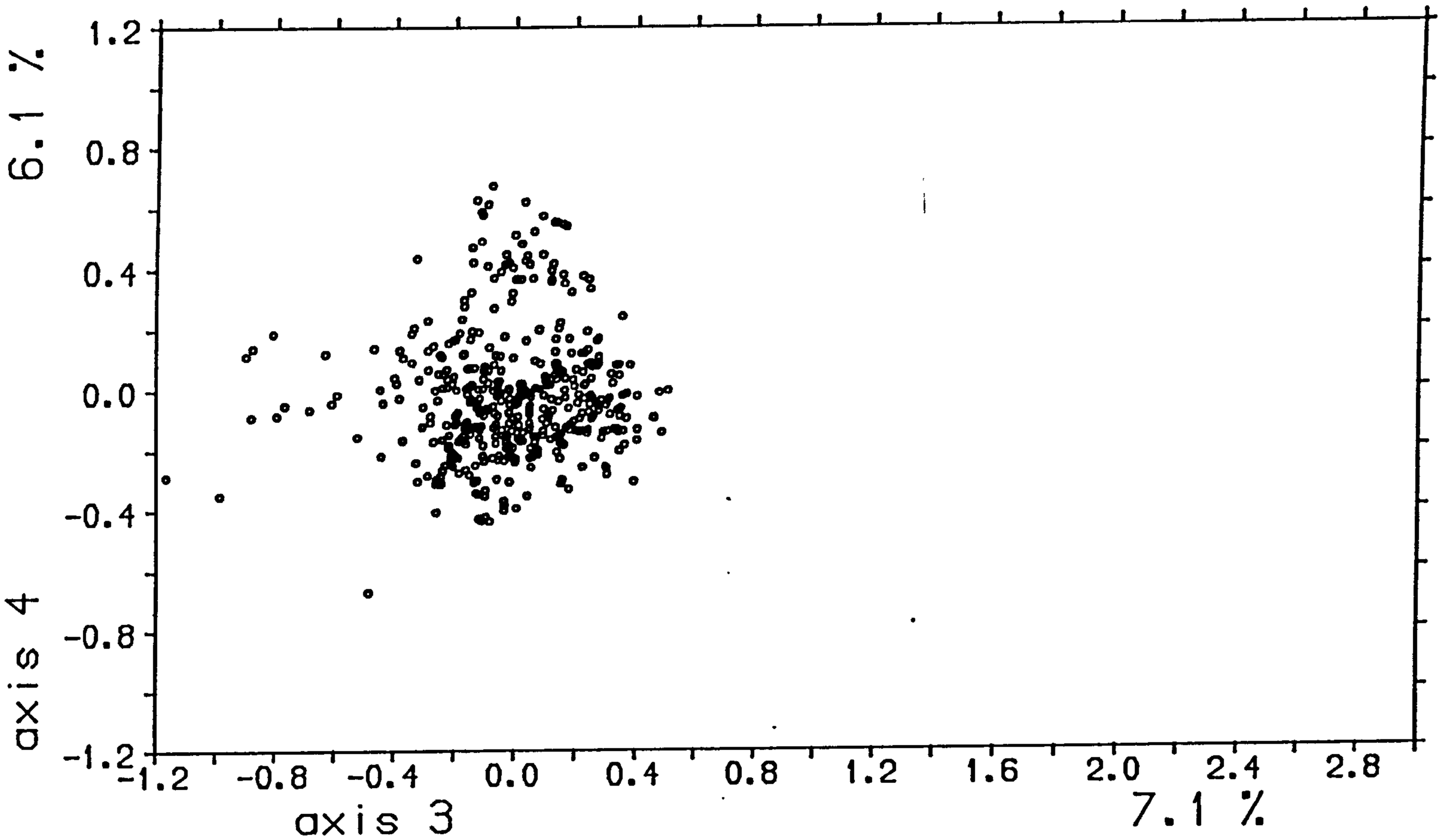

ALL 428 mRNAs
10.4 %
1.2
0.8
0.4
0.0
-0.4
-0.8
-1.2
axis 2
-1.2
-0.8
-0.4
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
axis 1
20.0 %

Opposite page 60

Figure 2.05

All 428 genes involved in the correspondence analysis plotted according to their co-ordinates on principal axes three and four.

ALL 428 mRNAs



Opposite page 61

Figure 2.06

The sixty-one sense codons of the universal genetic code plotted according to their co-ordinates on principal axes one and two. Compare with Figure 2.04 (the 428 genes plotted on the same 2D plot).

61 CODONS

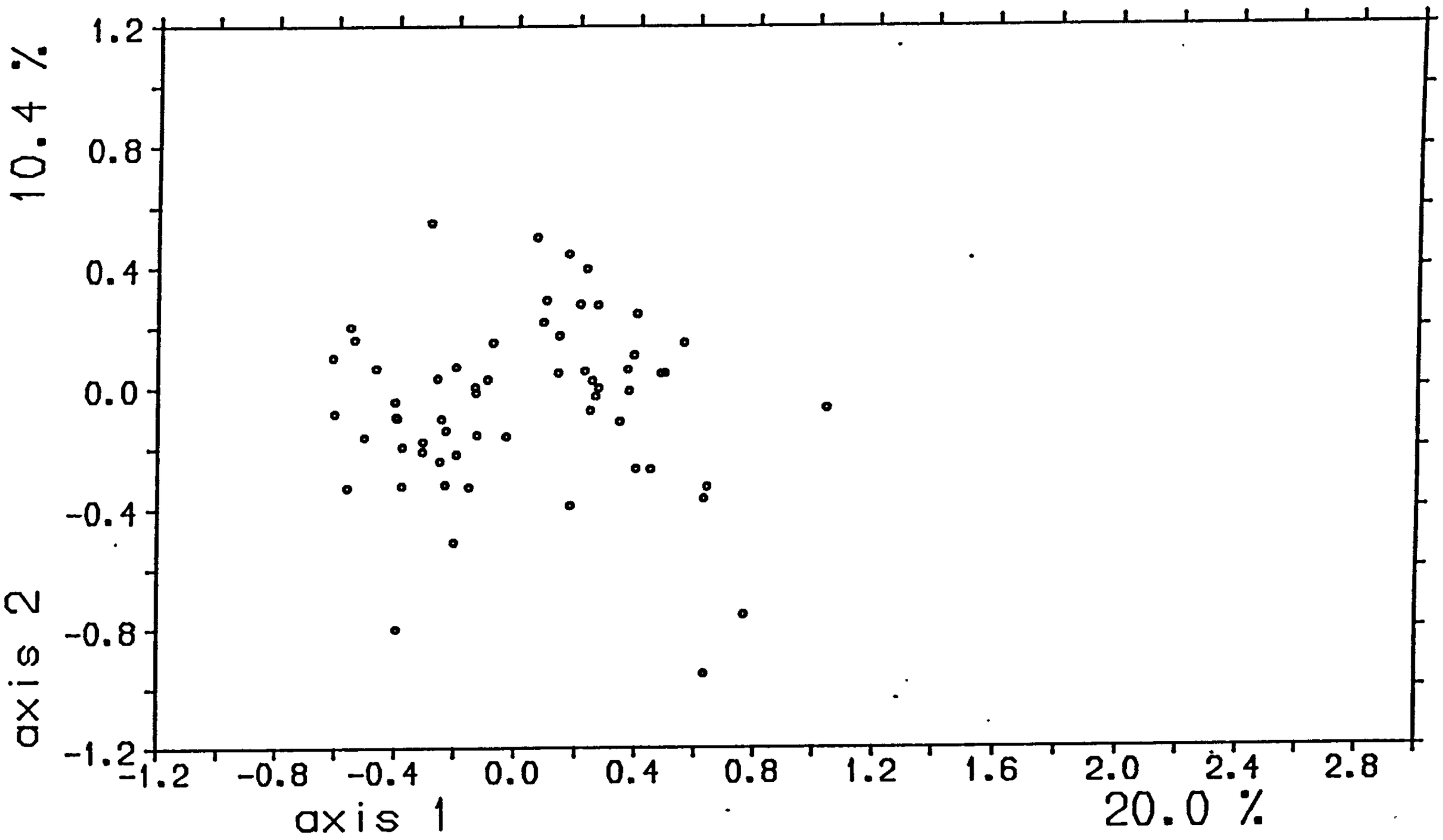

10.4 %
1.2
0.8
0.4
0.0
-0.4
-0.8
-1.2
axis 2
-1.2
-0.8
-0.4
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
axis 1
20.0 %

Figure 2.07

The sixty-one sense codons of the universal genetic code plotted according to their co-ordinates on principal axes three and four. Compare with Figure 2.05 (the 428 genes plotted on the same 2D plot).

61 CODONS

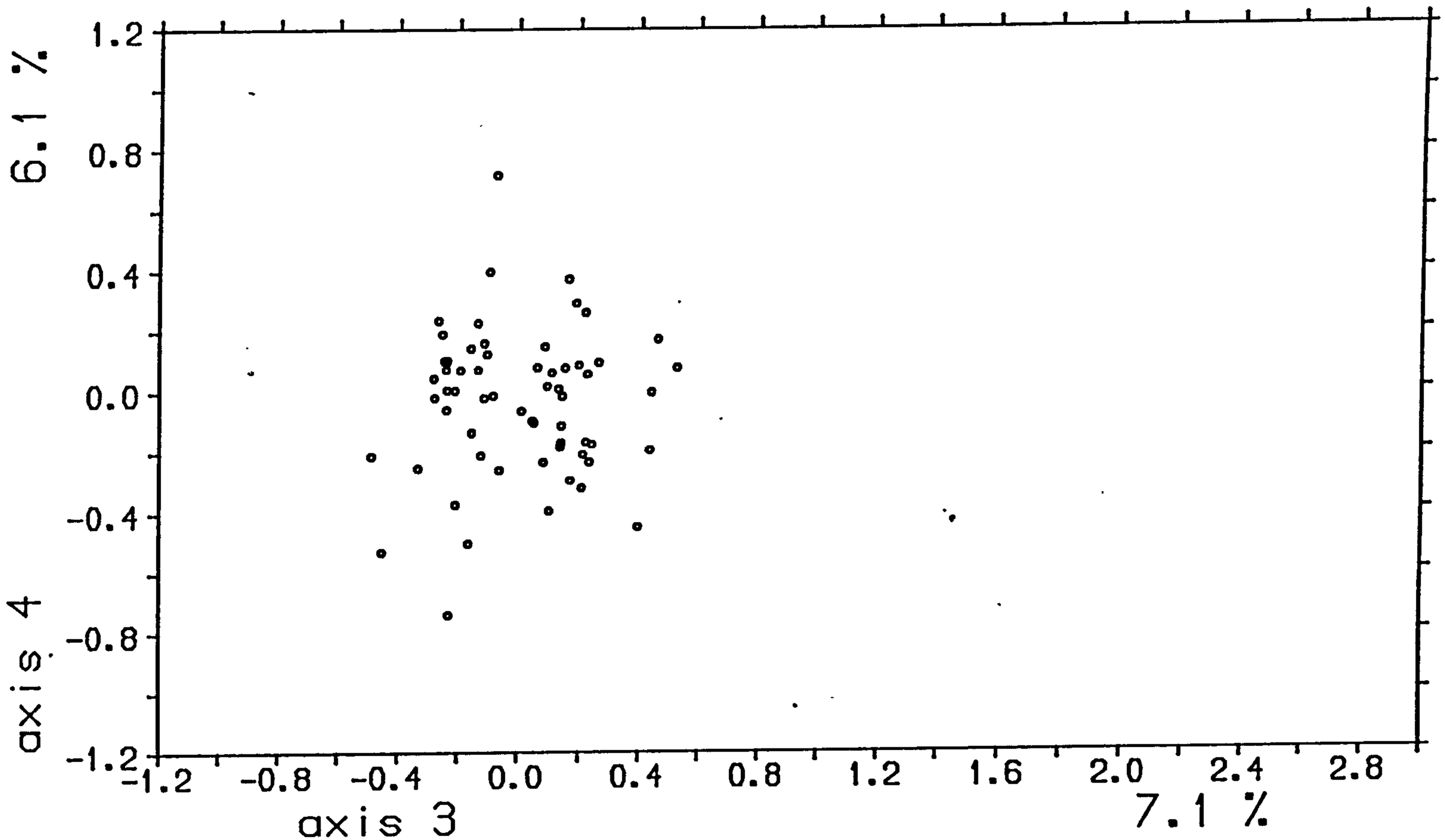

6.1 %
1.2
0.8
0.4
0.0
-0.4
-0.8
-1.2
axis 4
-1.2
-0.8
-0.4
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
axis 3
7.1 %

Table 2.5

The co-ordinates of the sixty-one sense codons w.r.t. the first four principal axes (see note on page 65). This data is plotted in Figures 2.06 and 2.07.

Table 2.5: Co-ordinates of the 61 Codons w.r.t. the New Axes.

	codon/a.a		PA:1	PA:2	PA:3	PA:4
	UUU/Phe	1	0.340	-0.111	0.241	-0.173
e y	UUC/Phe	2	-0.031	-0.158	-0.135	0.229
T7	UUA/Leu	3	1.036	-0.068	0.440	-0.003
y	UUG/Leu	4	0.140	0.173	-0.332	-0.250
	CUU/Leu	5	0.243	-0.075	0.048	-0.094
	CUC/Leu	6	-0.204	-0.509	-0.191	0.074
T7	CUA/Leu	7	0.634	-0.951	-0.073	0.717
e	CUG/Leu	8	-0.612	0.103	0.265	0.096
T7 y	AUU/Ile	9	0.477	0.046	0.200	0.088
T7 e y	AUC/Ile	10	-0.134	-0.014	-0.095	0.397
T7	AUA/Ile	11	0.766	-0.753	0.221	0.260
T7 - -	AUG/Met	12	-0.077	0.151	0.013	-0.063
e y	GUU/Val	13	0.169	0.444	-0.101	0.126
y	GUC/Val	14	-0.232	-0.138	-0.210	0.006
T7 e	GUA/Val	15	0.388	0.109	0.190	0.292
	GUG/Val	16	-0.505	-0.161	0.084	-0.232
e y	UCU/Ser	17	0.266	0.274	-0.235	0.008
e y	UCC/Ser	18	-0.247	-0.101	-0.235	0.105
	UCA/Ser	19	0.639	-0.328	0.134	0.010
T7	UCG/Ser	20	-0.379	-0.193	0.235	-0.232
	CCU/Pro	21	0.271	-0.000	-0.153	-0.132
	CCC/Pro	22	-0.397	-0.800	-0.277	-0.016
y	CCA/Pro	23	0.398	-0.266	-0.239	0.076
T7 e	CCG/Pro	24	-0.555	0.203	0.462	0.171
e y	ACU/Thr	25	0.397	0.245	-0.237	-0.058
e y	ACC/Thr	26	-0.309	-0.176	-0.112	0.161
	ACA/Thr	27	0.629	-0.366	0.097	0.019
T7	ACG/Thr	28	-0.400	-0.094	0.435	-0.194
T7 e y	GCU/Ala	29	0.229	0.395	-0.280	0.047
T7 y	GCC/Ala	30	-0.380	-0.323	-0.081	-0.011
T7 e	GCA/Ala	31	0.139	0.050	0.155	0.080
T7	GCG/Ala	32	-0.541	0.161	0.526	0.078

Table 2.5 contd.: Co-ordinates of the 61 codons w.r.t. the New Axes.

	codon/a.a		PA:1	PA:2	PA:3	PA:4
T7	UAU/Tyr	33	0.367	0.060	0.223	-0.165
T7 e y	UAC/Tyr	34	-0.129	-0.153	-0.245	0.105
	...	..	.....	.....	.....	.....
	...	..	.....	.....	.....	.....
T7	CAU/His	35	0.261	-0.026	0.214	-0.207
e y	CAC/His	36	-0.196	-0.218	-0.252	0.192
y	CAA/Gln	37	0.372	-0.009	-0.110	-0.019
e	CAG/Gln	38	-0.466	0.067	0.144	-0.112
T7	AAU/Asn	39	0.492	0.048	0.210	-0.318
e y	AAC/Asn	40	-0.137	0.005	-0.157	0.145
e	AAA/Lys	41	0.207	0.277	0.146	-0.014
y	AAG/Lys	42	-0.200	0.071	-0.487	-0.210
T7	GAU/Asp	43	0.087	0.218	0.140	-0.182
y	GAC/Asp	44	-0.262	0.033	-0.131	0.075
e y	GAA/Glu	45	0.096	0.290	0.064	0.082
	GAG/Glu	46	-0.396	-0.098	-0.120	-0.209
y	UGU/Cys	47	0.250	0.025	-0.163	-0.500
T7	UGC/Cys	48	-0.233	-0.319	-0.206	-0.370
	...	..	.....	.....	.....	.....
- -	UGG/Trp	49	-0.093	0.030	-0.060	-0.256
T7 e	CGU/Arg	50	-0.289	0.549	0.165	0.371
T7 e	CGC/Arg	51	-0.605	-0.084	0.229	0.058
T7	CGA/Arg	52	0.181	-0.388	0.112	0.064
T7	CGG/Arg	53	-0.561	-0.328	0.397	-0.449
	AGU/Ser	54	0.225	0.056	0.103	-0.392
T7 e	AGC/Ser	55	-0.310	-0.209	0.142	-0.167
y	AGA/Arg	56	0.555	0.148	-0.452	-0.527
	AGG/Arg	57	-0.155	-0.327	-0.228	-0.739
e y	GGU/Gly	58	0.064	0.500	-0.265	0.237
T7 e	GGC/Gly	59	-0.404	-0.044	0.088	0.149
	GGA/Gly	60	0.446	-0.269	0.053	-0.101
	GGG/Gly	61	-0.251	-0.240	0.173	-0.293

Table 2.5 contd.: Co-ordinates of the 61 Codons w.r.t. the New Axes.

Abbreviations:

PA = principal axis.

T7 = Bacteriophage T7 codons that partly comprise palindromes.

e = *E.coli* optimal codons.

y = Yeast optimal codons.

Information from Gene Plots – Principal Axes 1 and 2.

The plots of the 428 genes contain much detailed information on codon usage patterns between and within species. Relationships between host and virus/bacteriophage and between organelle and nuclear genes, among other relationships, are likely to be apparent in the displays. The object of this correspondence analysis was to look for general patterns of codon usage and this involved looking in some detail at some aspects of a particular species codon usage patterns. However, it is not the intention to dwell on the details of individual codon usage patterns *per se*

The five plots of principal axis 1 versus axis 2 for the five groups are shown in Figures 2.08 to 2.12. Eukaryotic Nuclear Genes (Figure 2.08) consist of nine taxonomic groups, six of whom are species. Most of the scatter is on the first principal axis (tentatively labelled "third position G+C content" above). The nine taxonomic groups appear to differ in their mean values for this principal axis: the vertebrate species (man, chicken, pisces sp., and *X.laevis*) have low values (=high G+C), whereas Yeast have high values (=low G+C). There is considerable intra-specific variation w.r.t. this first axis also, although there is little overlap between the vertebrates and Yeast. The Eukaryotic Viral Genes (Figure 2.09) show a similar pattern to the multicellular Eukaryotic Nuclear Genes. Note the considerable intra-specific variation in Adenovirus 2 w.r.t. the first axis.

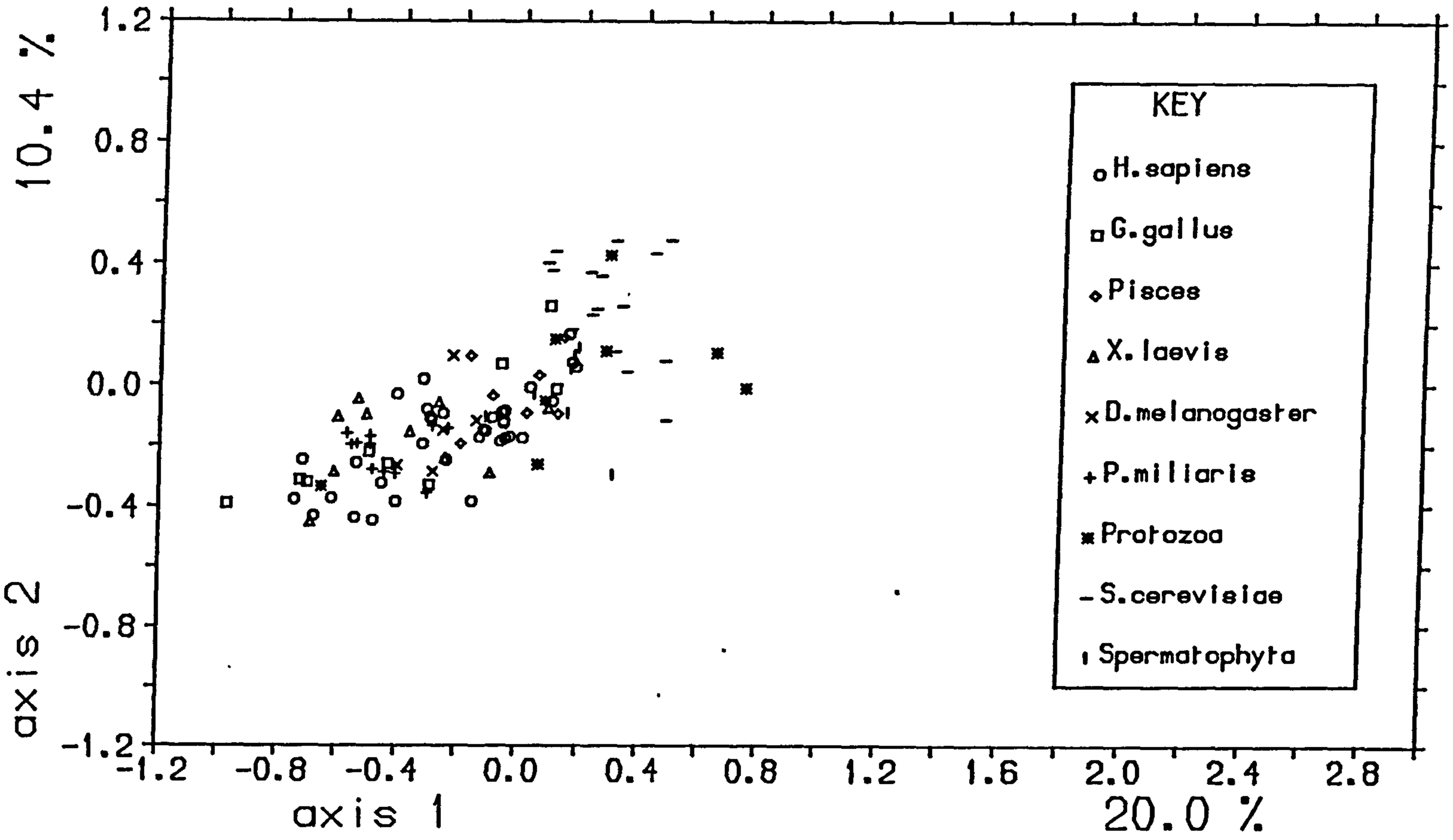
The Eukaryotic Organelle Genes (Figure 2.10) show a different pattern of codon usage from nuclear genes except in the case of chloroplast and

Opposite page 66

Figure 2.08

Part of the the output from the single correspondence analysis of the codon usage patterns of 428 genes. The nine Eukaryotic Nuclear subgroups are described in more detail in table 2.4.

EUKARYOTIC NUCLEAR GENES

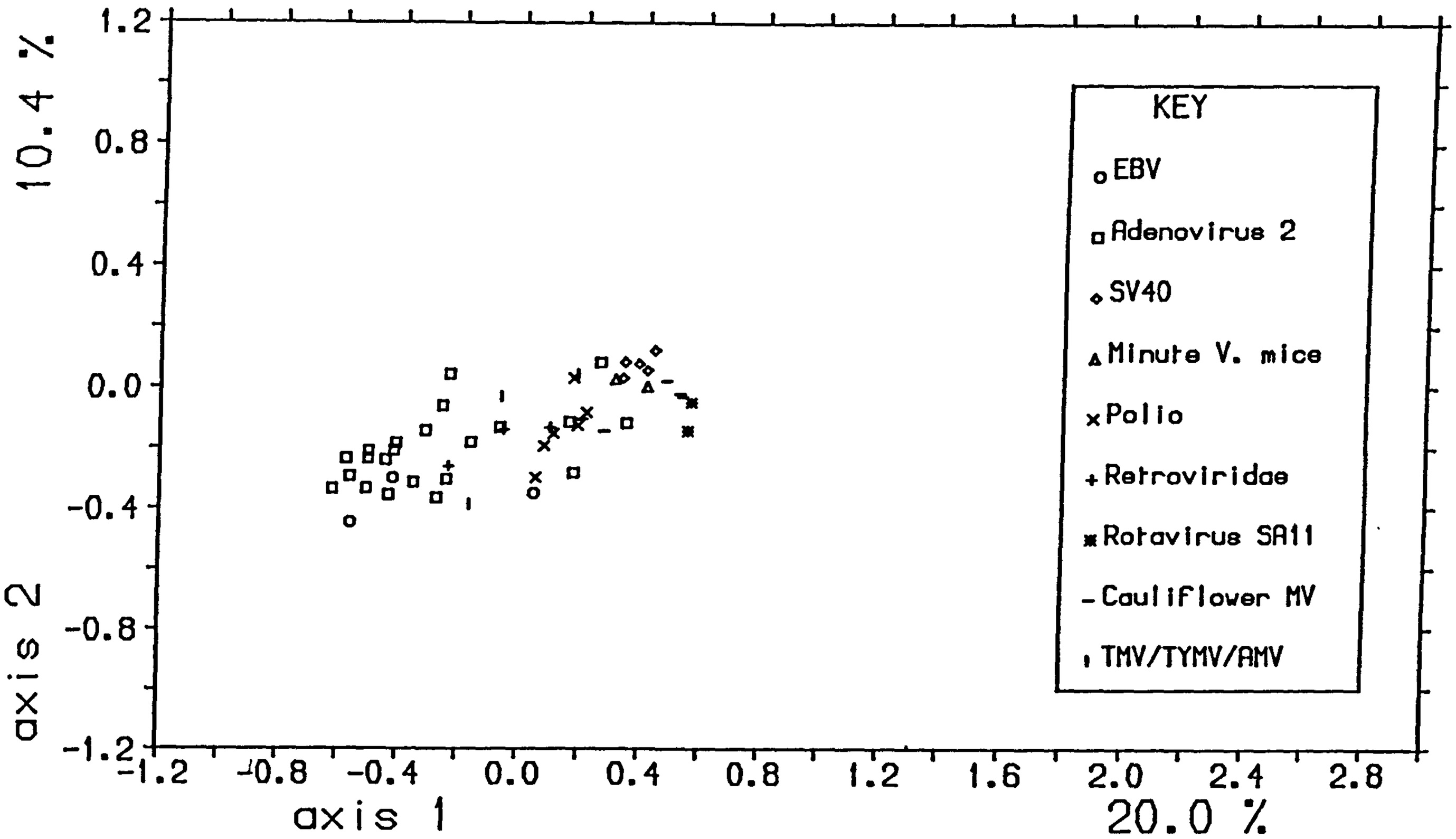


Opposite page 67

Figure 2.09

Part of the the output from the single correspondence analysis of the codon usage patterns of 428 genes. The nine Eukaryotic Viral subgroups are described in more detail in table 2.4.

EUKARYOTIC VIRAL GENES

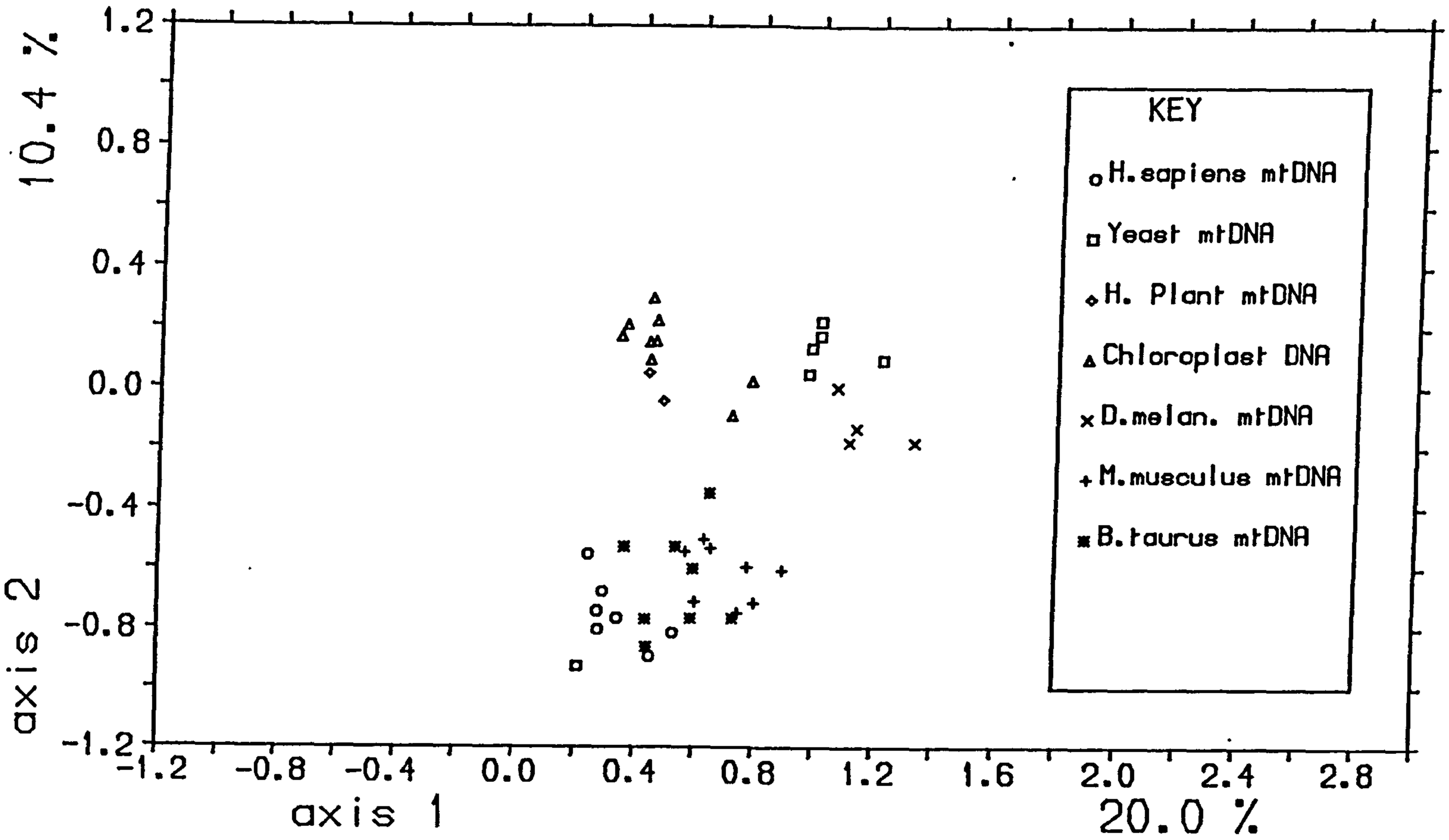


Opposite page 68

Figure 2.10

Part of the the output from the single correspondence analysis of the codon usage patterns of 428 genes. The seven Eukaryotic Organelle subgroups are described in more detail in table 2.4.

EUKARYOTIC ORGANELLE GENES

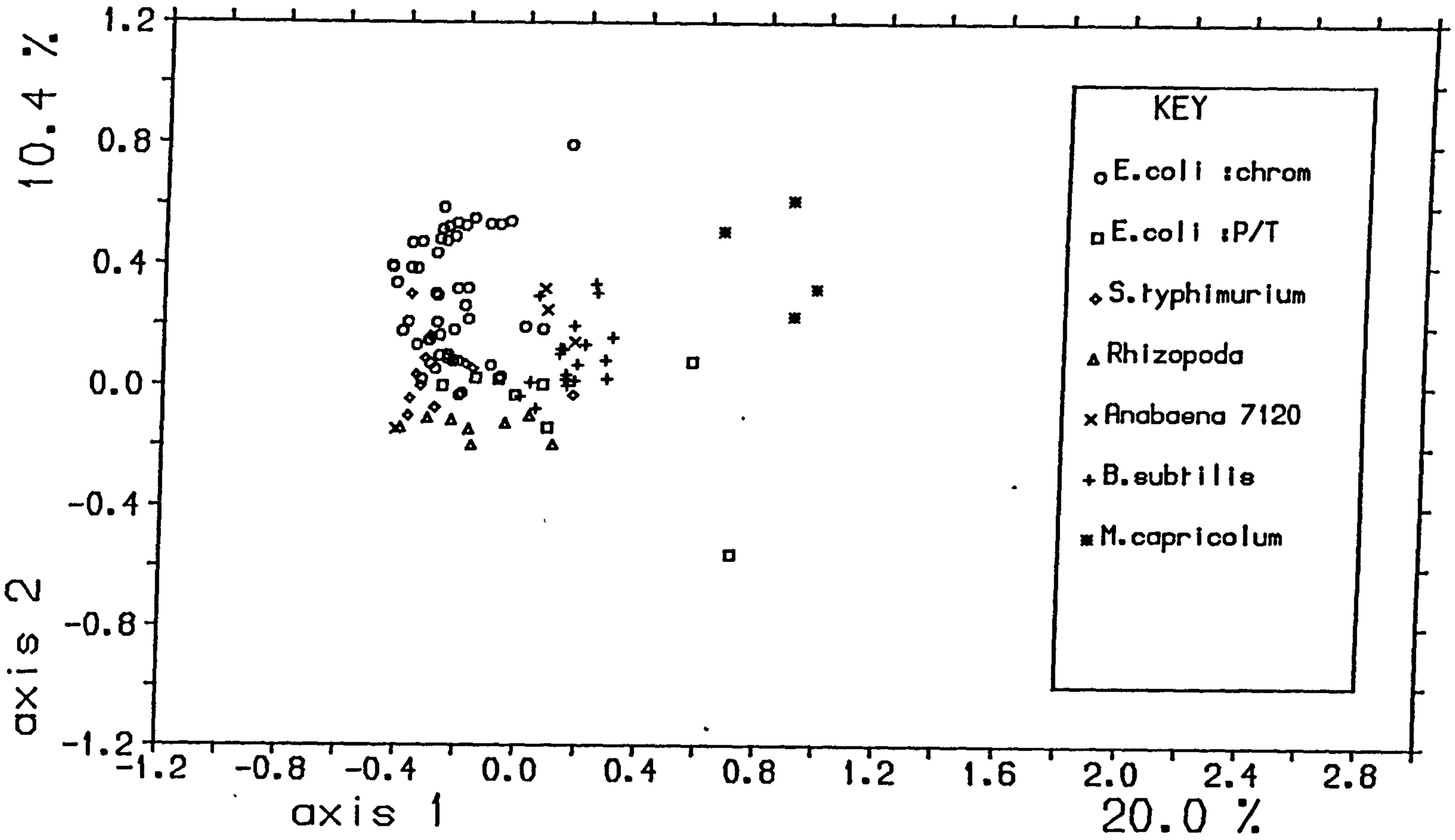


Opposite page 69

Figure 2.11

Part of the the output from the single correspondence analysis of the codon usage patterns of 428 genes. The seven Prokaryotic subgroups are described in more detail in table 2.4.

PROKARYOTIC GENES

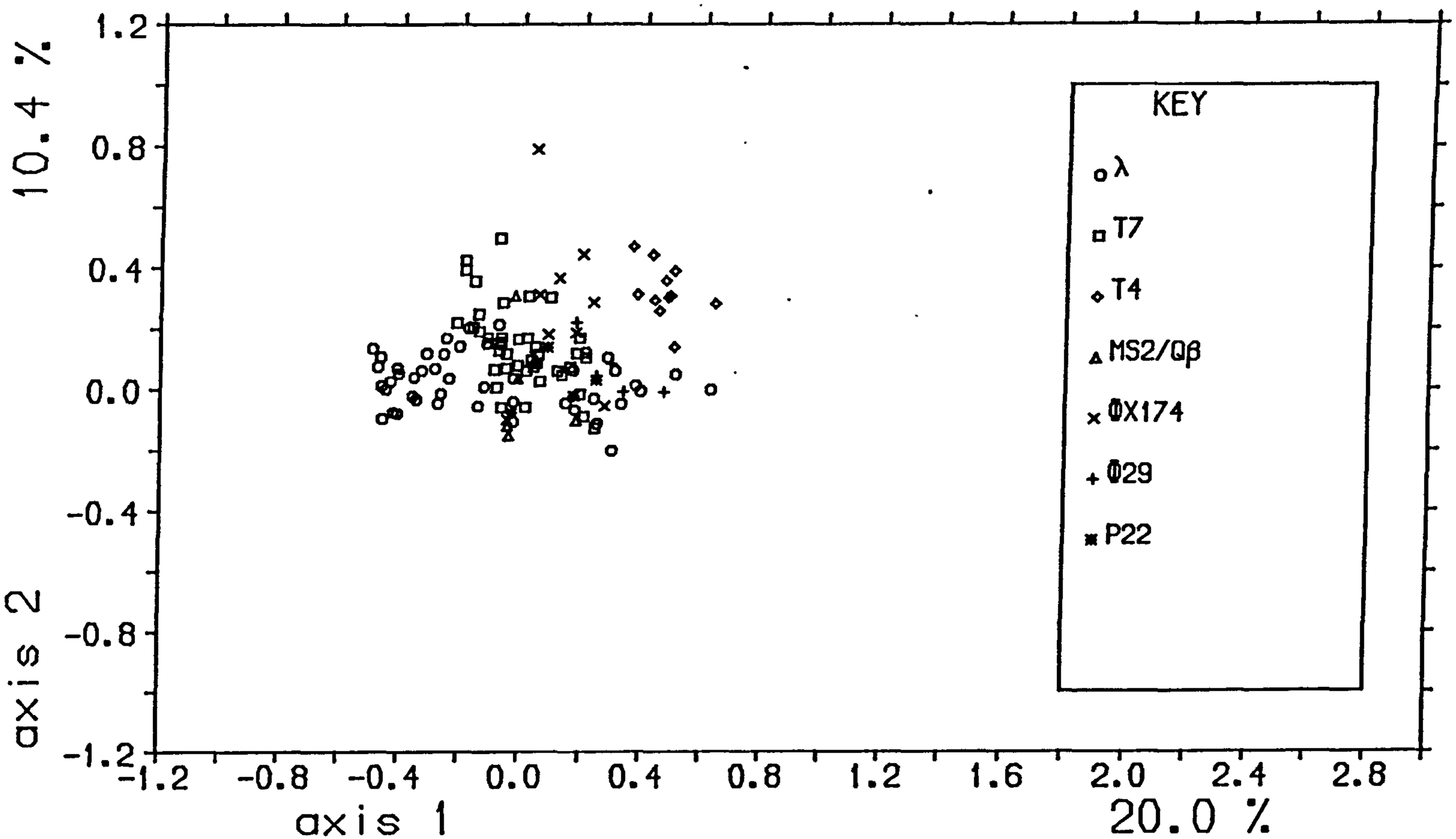


Opposite page 70

Figure 2.12

Part of the the output from the single correspondence analysis of the codon usage patterns of 428 genes. The seven Bacteriophage subgroups are described in more detail in table 2.4.

BACTERIOPHAGE GENES



mitochondria of higher plants. The mitochondria of *D.melanogaster* and Yeast show very high values on principal axis 1 due to their low G+C content. Genes from the three mammalian mitochondria have very low values on the second principal axis. The positioning of these five mitochondria at the extremes of the plot suggests that they had a considerable effect in determining the positions of the first two principal axes.

The Prokaryotic Genes (Figure 2.11) consist of a wide range of bacterial species. The *E.coli* genes have been split into two categories: chromosomal genes and plasmid-borne/transposon genes. There is a close similarity between *E.coli* and *S.typhimurium* codon usage. The various bacterial taxonomic groups show variation in their mean value for axes 1 (tentatively third- position G+C content), whereas most of the within-species variation is along axis 2. Since the first two principal axes are orthogonal to each other, this suggests that this within-species variation is independent of G+C content. The intra-specific variation in *E.coli* genes is related to the level of gene expression: highly-expressed genes (e.g. ribosomal proteins) have high axis 2 scores.

Most bacteriophages are entirely dependent on their bacterial hosts for tRNAs to decode their messenger RNA. The eight Bacteriophages (Figure 2.12 and 2.17) are dominated by coliphages: ϕ29 and P22 are the exceptions. MS2 and Qβ are closely related and have been pooled due to the small sample size. The coliphages show much variation, both in their mean value and their within- species variation, along axis 1. Bacteriophages T7 and ϕX174 show the most similar pattern to *E.coli*. Bacteriophage λ shows a wide range of G+C content in agreement with the known G+C diversity of its segmented genome. The codon usage differences between the lytic T7 phage and the lysogenic λ phage are obvious: the lytic phage appears to resemble the host codon usage more than does the lysogenic phage. This relationship has also been noted by Grantham *et al.* (1985). Bacteriophage T4 has a small complement of 8 tRNAs and is not totally dependent on its host: this may explain why its codon usage is not very similar to that of *E.coli*.

Information from Gene Plots – Principal Axes 3 and 4.

The plane defined by the third and fourth principal axes displays 13.2% of the Total Inertia. Consideration of the plot of all 428 genes on axes 3 and 4 (Figure 2.05), reveals that comparatively few genes show scatter on axis 3. Comparison of Figure 2.05 with the five sub-plots (Figures 2.13 to 2.17) reveals that these genes belong to Yeast and Protozoa. Care must be exercised in drawing conclusions from the Protozoan genes due to the large evolutionary distances between these genes. The remarkable intra-specific variation shown by the 16 Yeast genes is related to the level of gene expression: the genes with large negative scores on axes 3 are all highly expressed. The (very) limited Protozoan data suggest that this pattern of codon usage may be common to unicellular eukaryotes.

Mammalian mitochondrial genes have the highest scores on axes 4, closely followed by *E.coli* highly expressed genes. This can be seen from Figures 2.15 and 2.17 respectively. Little scatter is seen on the other plots for this axis. As noted above, codons AGG and AGA have high negative scores for axes 4. These two codons are also rarely used in *E.coli* highly-expressed genes (data not shown). This axis may be showing common features of *E.coli* and mammalian mitochondrial genes. Note that these two groups were at opposite ends of axis 2 (see Figures 2.10 and 2.12).

2.7.3. Supplementary Profiles.

The interpretation of the correspondence analysis output so far has followed the approach of Greenacre & Vrba (1984). An additional aid to interpreting the displays is outlined in Greenacre (1984): the plotting of reference points on the plots. These "supplementary profiles", (SPPs), are plotted using the transition formula given in equation (2.12) in section 2.4.1. This facility was used to carry out a preliminary study of the usefulness of two classes of codon usage models:

1. Models based on mono- and di-nucleotide composition. These will be termed "simple nucleotide models".

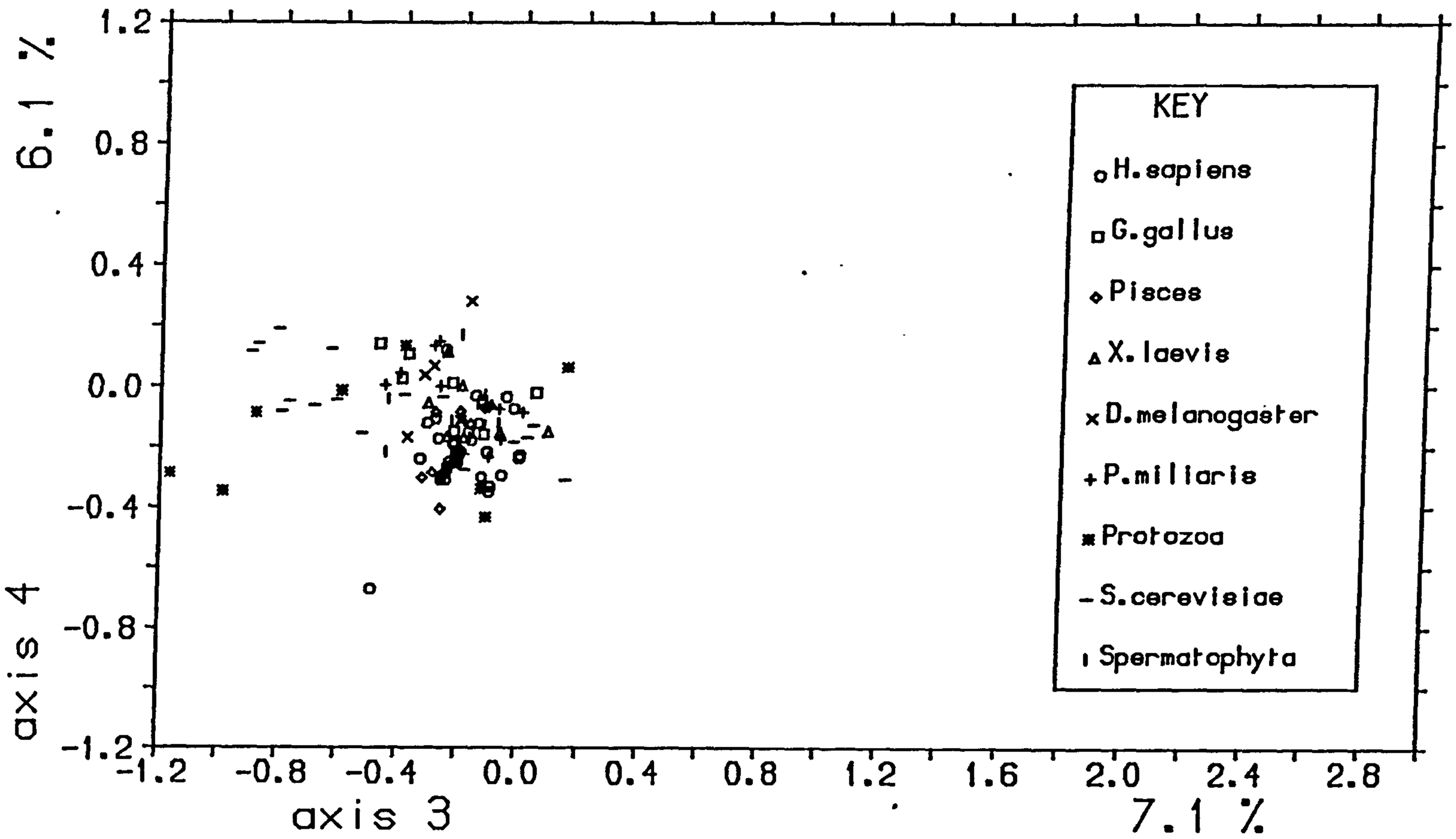
2. Models based on tri-nucleotide composition. These will be termed "codon models".

Opposite page 73

Figure 2.13

Part of the the output from the single correspondence analysis of the codon usage patterns of 428 genes.

EUKARYOTIC NUCLEAR GENES

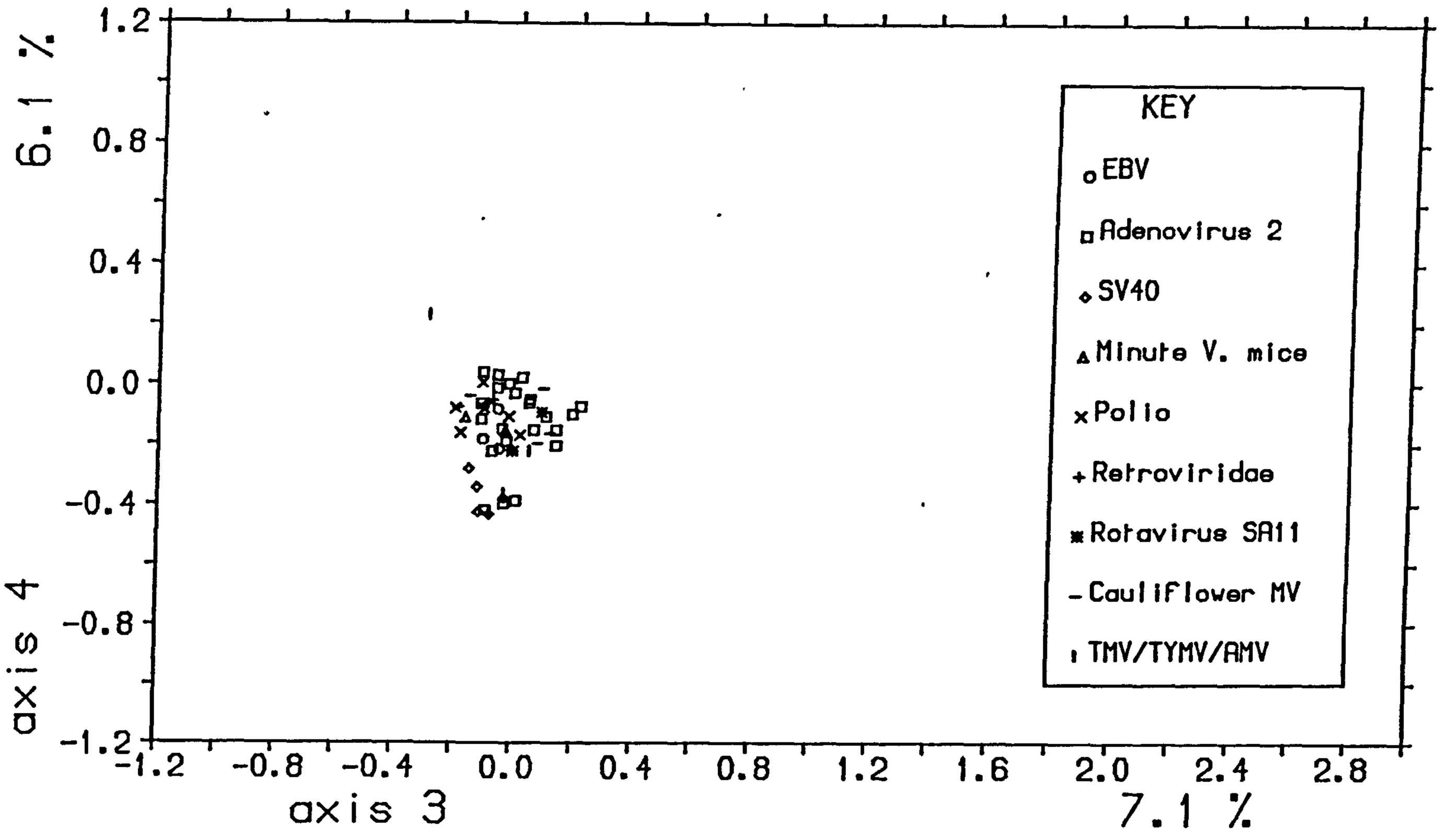


Opposite page 74

Figure 2.14

Part of the the output from the single correspondence analysis of the codon usage patterns of 428 genes.

EUKARYOTIC VIRAL GENES

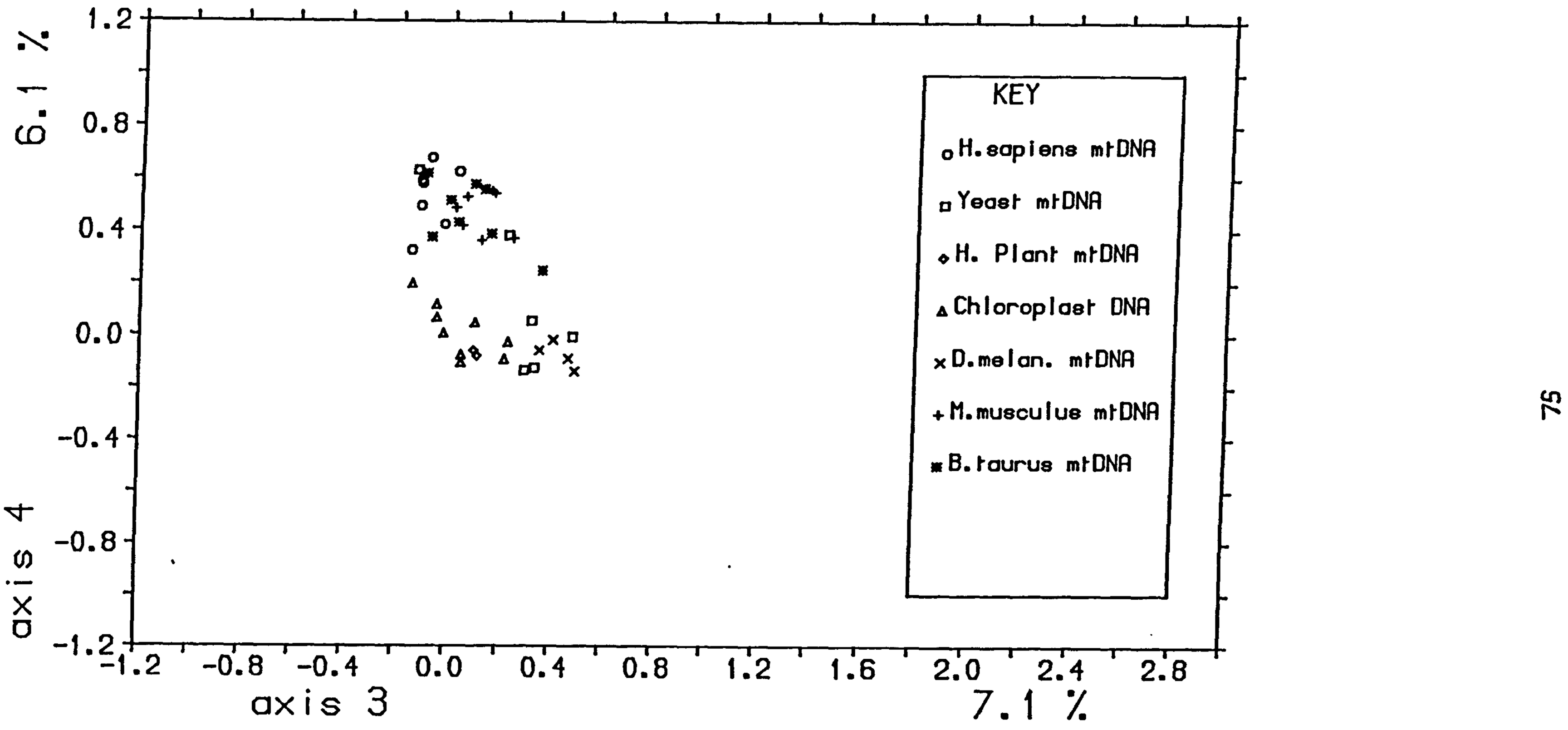


Opposite page 75

Figure 2.15

Part of the the output from the single correspondence analysis of the codon usage patterns of 428 genes.

EUKARYOTIC ORGANELLE GENES

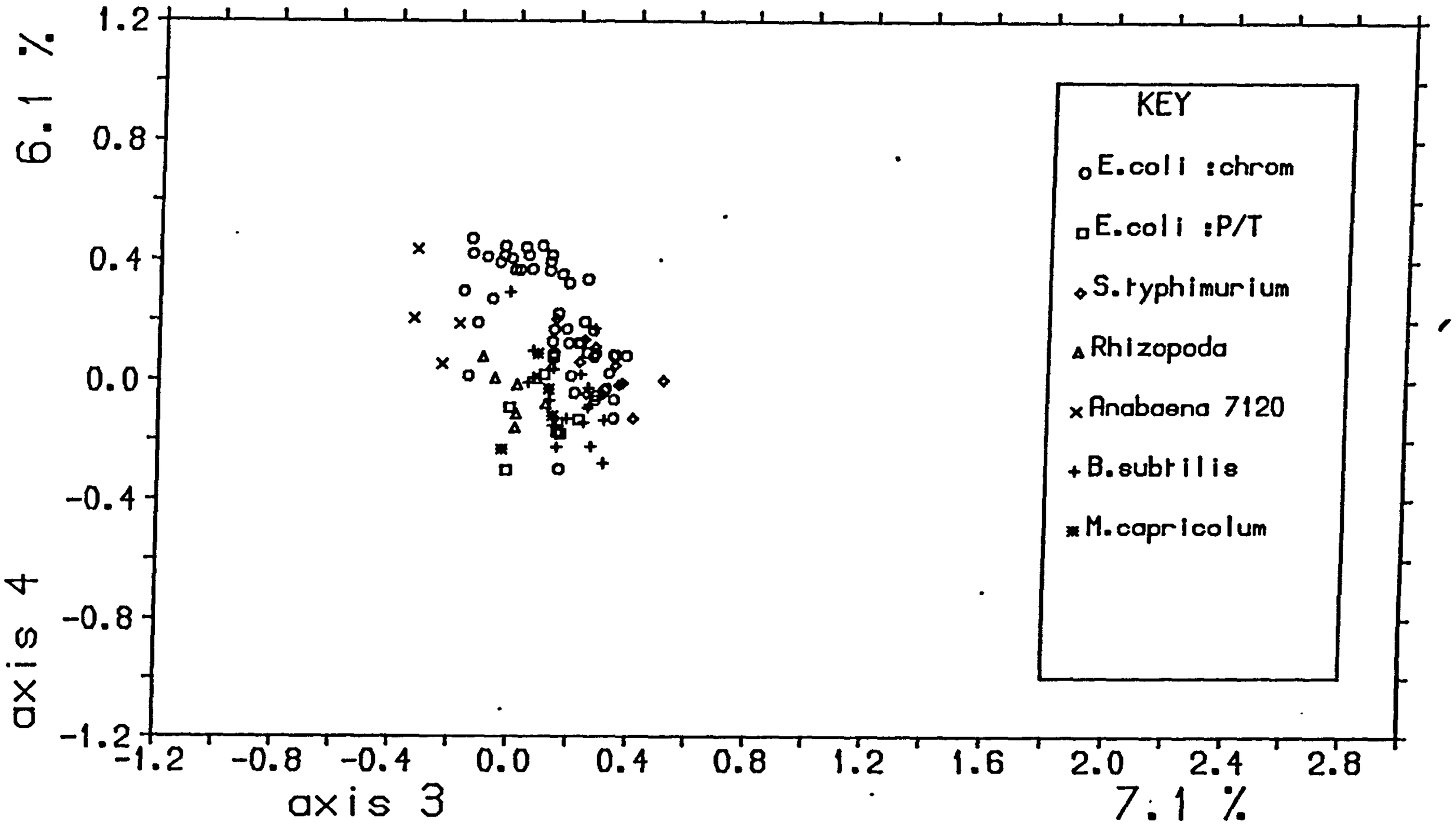


Opposite page 76

Figure 2.16

Part of the the output from the single correspondence analysis of the codon usage patterns of 428 genes.

PROKARYOTIC GENES

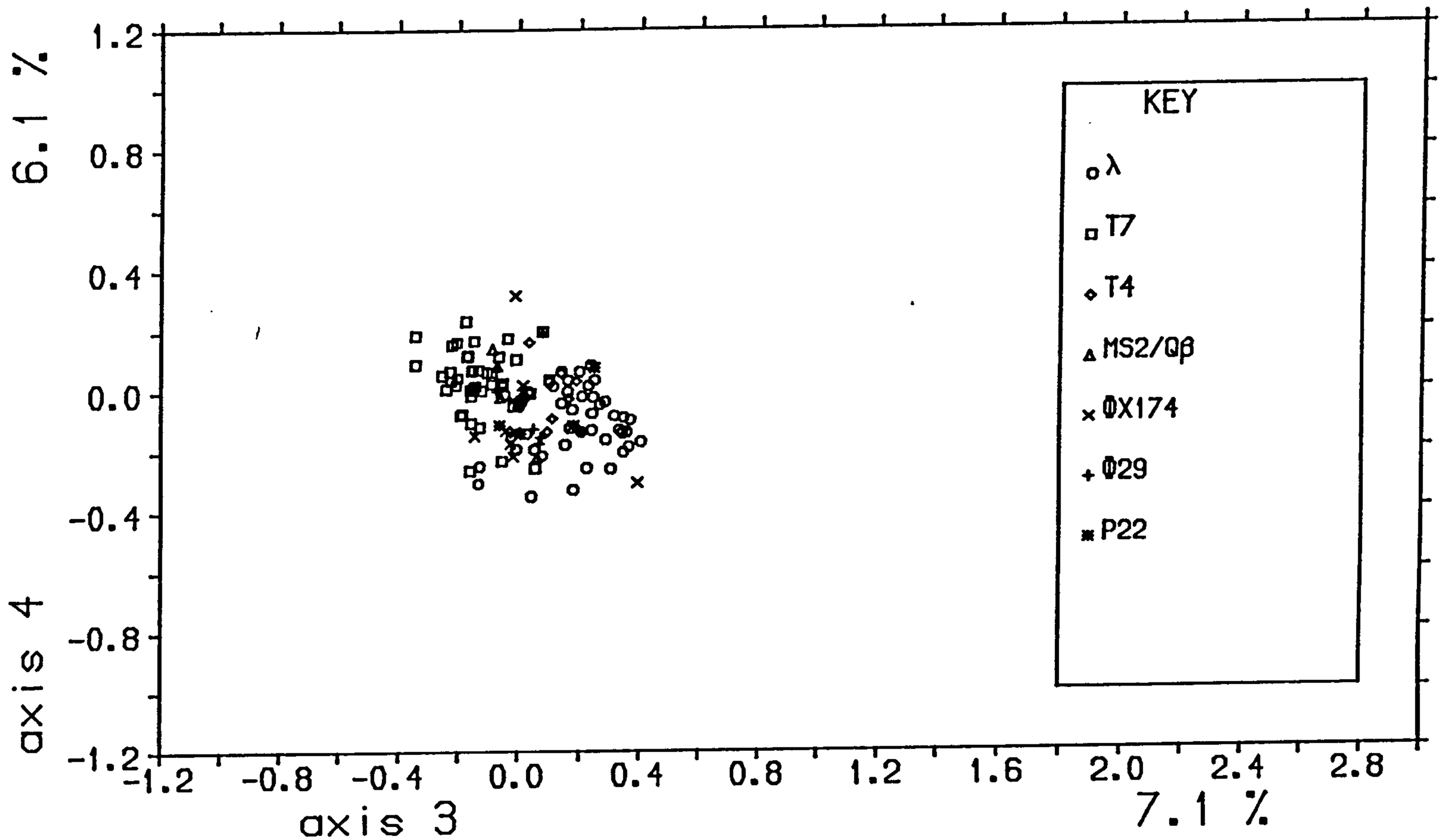


Opposite page 77

Figure 2.17

Part of the the output from the single correspondence analysis of the codon usage patterns of 428 genes.

BACTERIOPHAGE GENES



Seven types of simple nucleotide model were examined. Using the amino-acid composition of the centroid codon usage pattern (the overall average), the third positions of the 61 sense codons were varied to examine:

- The effect of varying T (while keeping the relative frequencies of the other three nucleotides equal).
- The effect of varying C.
- The effect of varying A.
- The effect of varying G.

Di-nucleotide composition models were studied by carrying out a similar analysis using the three pairs of di-nucleotides:

- A+T content versus G+C content.
- A+C content versus G+T content.
- A+G content versus T+C content.

The relative frequencies of the nucleotides composing each dinucleotide pair were kept constant.

The second class of codon usage models, the codon models, were not simply related to mono- or di-nucleotide frequencies. Instead the abundance of a particular subset of codons was of interest. Four comparisons were carried out:

- A codon usage pattern where codons containing the CpG dinucleotide (i.e. CGN and NCG codons) are either at a maximum or minimum.
- A codon usage pattern where codons containing those codons of intermediate energy (according to Grosjean & Fiers 1982) are at a maximum or minimum. This model will be labelled "PB" to denote pyrimidine bias.
- A codon usage pattern where codons containing one

substitution event away from a stop codon are either at a maximum or minimum. This model will be labelled "PTC" to denote pre-termination codons.

- A codon usage pattern where codons of the form RNY are either at a maximum or minimum.

These four codon usage patterns were constrained to have the same amino-acid composition as the centroid.

Information from Simple Nucleotide Model SPP plots.

For each of the seven simple nucleotide models, a pair of points were generated by altering the third position nucleotide content of the average amino-acid composition in an appropriate manner. The two points for each model were then joined. Figures 2.18 and 2.19 show these seven lines plotted on the first four principal axes.

It is immediately apparent from Figure 2.18 that simple nucleotide models have the capability of explaining a large amount of the scatter on axes 1 and 2. The (G+C v A+T) supplementary profile line (SPL) spans axis 1 (high G+C is to the left of centre). Although this SPL is not exactly parallel to axis 1, it does appear to explain most of the scatter on the first two principal axes for the plots of the Eukaryotic Nuclear Genes (Figure 2.08) and the Eukaryotic Viral Genes (Figure 2.09). The range of this SPL also spans the total variation on axis 1 (see Figure 2.04).

The second principal axis is less well explained by the simple nucleotide SPLs. The codon usage bias of the three mammalian mitochondrial genomes is reasonably well described in terms of preferred usage of codons ending in -C and -A. The highly expressed *E.coli* genes reflect a bias in the usage of T, or T and G, in the third position but the scatter is not fully explained.

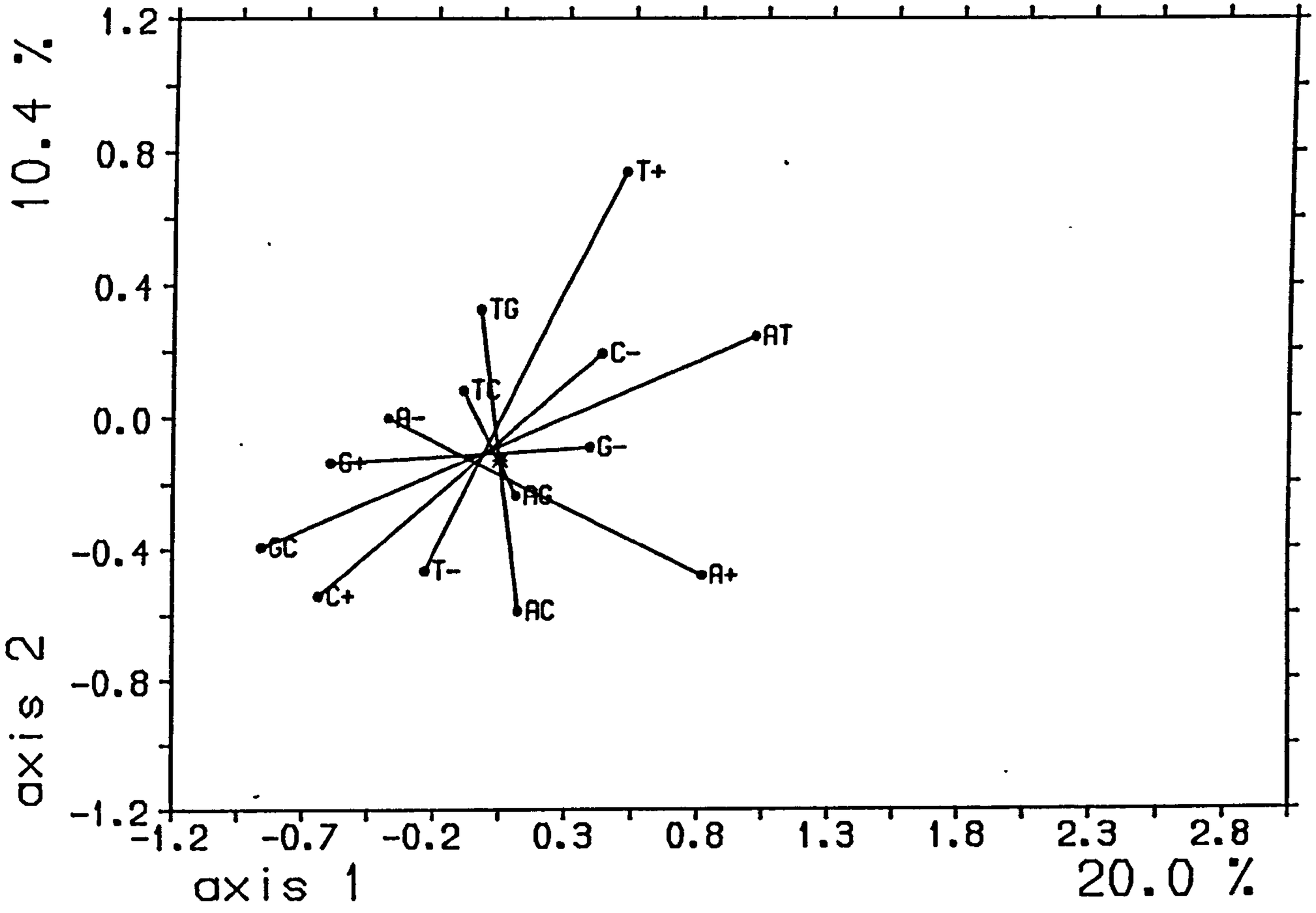
The scatter on the third principal axis is not well explained in terms of nucleotide composition, although part of the pattern is due to differences in A+G usage in the third position. A large part of the scatter on axis four appears due to G+T content differences.

Opposite page 80

Figure 2.18

Supplementary profiles for the simple nucleotide models (see page 79).

SPPs : SIMPLE NUCLEOTIDE MODELS

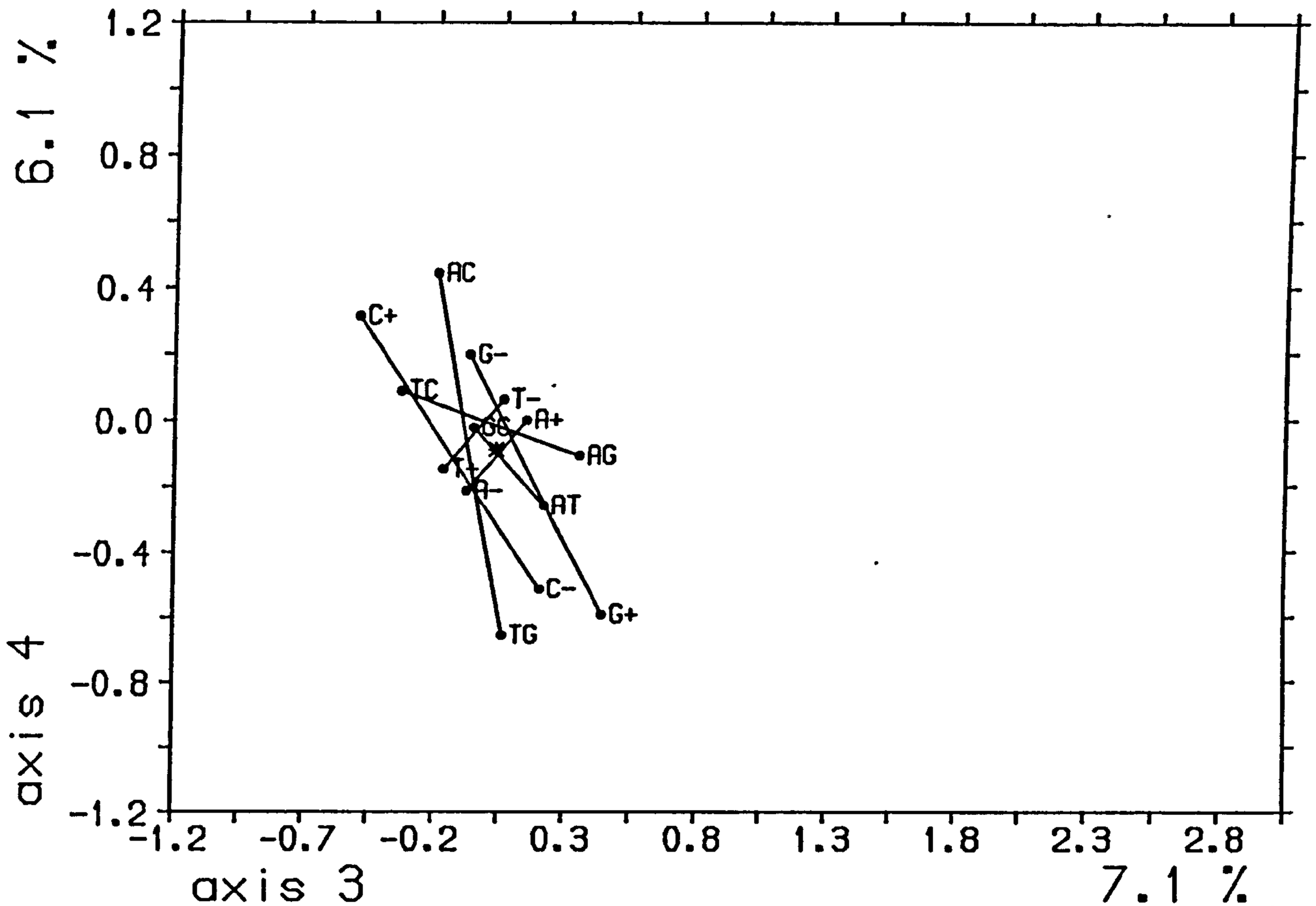


Opposite page 81

Figure 2.19

Supplementary profiles for the simple nucleotide models (see page 79).

SPPs : SIMPLE NUCLEOTIDE MODELS



Information from Codon Model SPP plots.

In comparison with the simple nucleotide SPP plots, the codon models SPP plots did not account for much of the observed scatter on the first four principal axes (see Figures 2.20 and 2.21). The association of the codon models with the positive ends of the first four principal axes is:

```
Principal Axis 1 (+):   associated with PTC(+), CpG(-).
Principal Axis 2 (+):   associated with RNY(+), PB(+).
Principal Axis 3 (+):   associated with CpG(+), RNY(-).
Principal Axis 4 (+):   associated with PB(+).
```

The first principal axis has been tentatively labelled as third position G+C content (high G+C values on the negative side of the axis). The position of low CpG on the positive side of axis 1 is thus not unexpected. CpG is low on the negative end of axis 3 (yeast highly expressed genes). PB is associated with the positive ends of axis 2 and axis 4. RNY is associated with high values on axis 2 and low values on axis 3.

2.7.4. The Codon Usage Pattern of the Centroid.

The correspondence analysis seeks to display the variation in the original data. The centroid of the cloud of genes (and the cloud of codons) is simply the average codon usage pattern of the set of genes studied (and also the relative lengths of the 428 genes). It therefore contains information on factors acting on all genes.

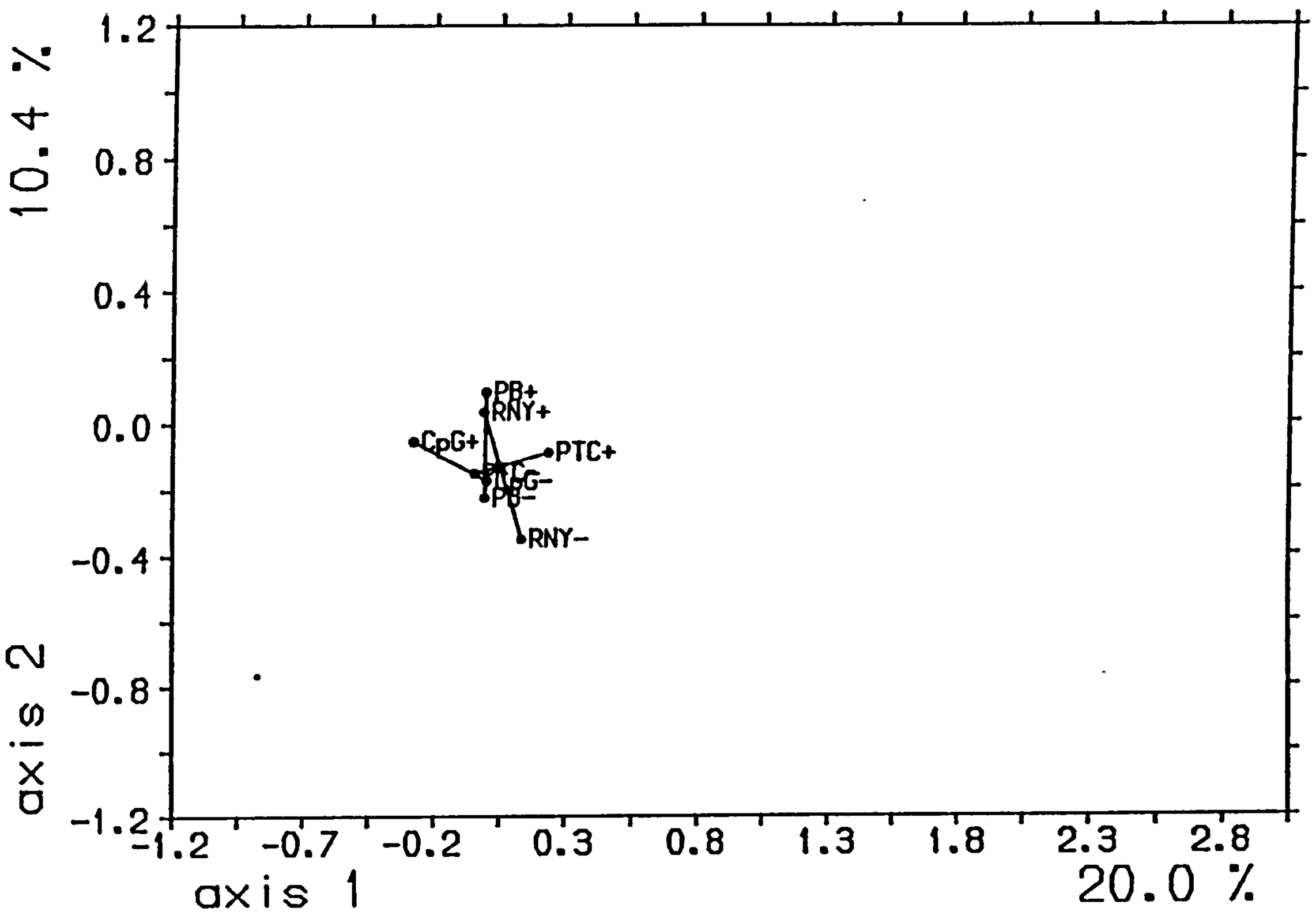
The codon usage pattern, the amino-acid composition, and the nucleotide composition of the centroid is shown in Table 2.6. The codon usage data has been standardised so that the entire table adds to 6100: thus if all codons were used equally the usage would be 100 for each codon. The total number of actual codons analysed was 119,824.

Opposite page 83

Figure 2.20

Supplementary profiles for the codon models (see page 82).

SPPs : CODON MODELS

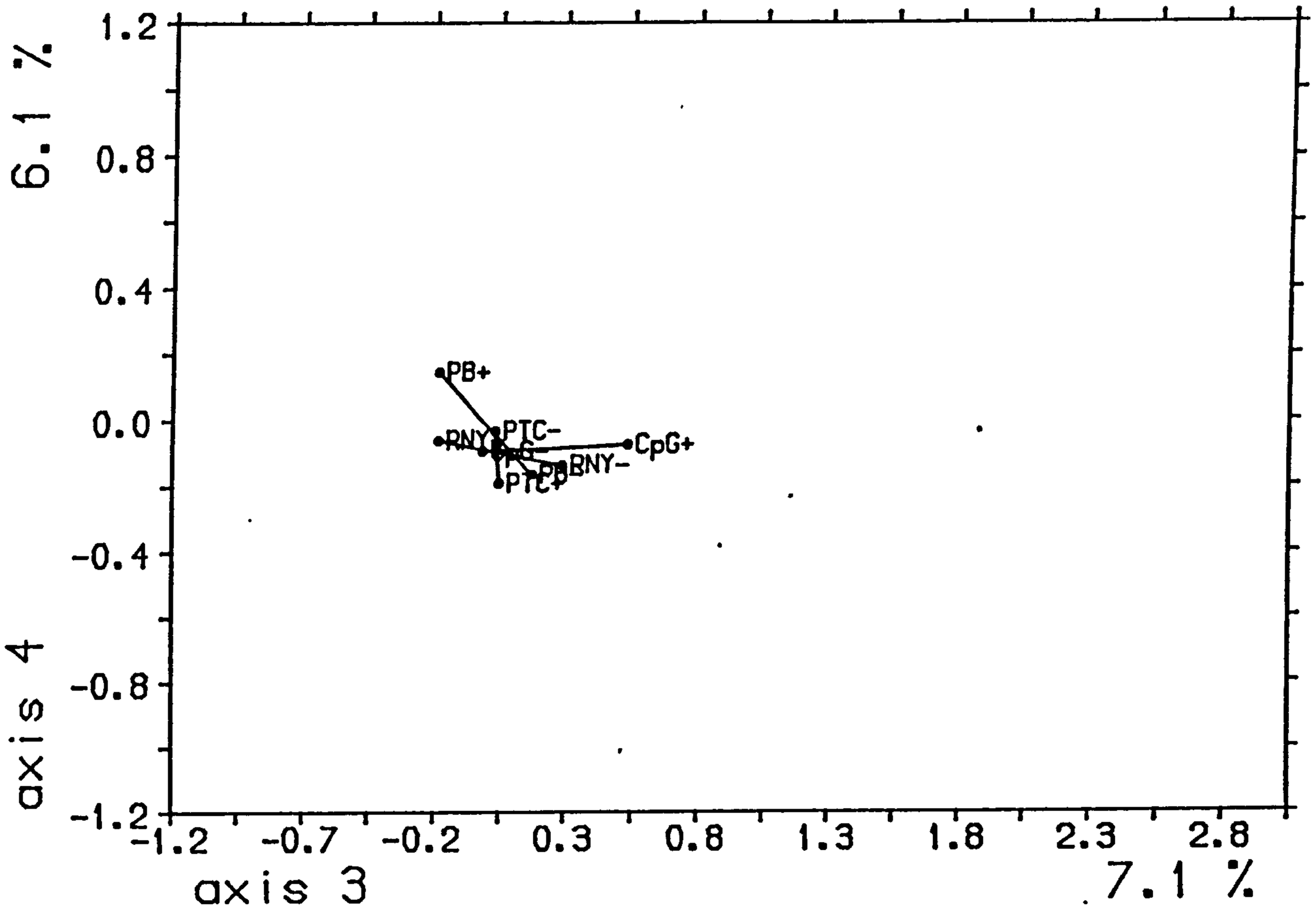


Opposite page 84

Figure 2.21

Supplementary profiles for the codon models (see page 82).

SPPs : CODON MODELS



Opposite page 85

Table 2.6a

The codon usage of the centroid (see page 82). Equal codon usage of all codons would have yielded a value of 100 for each codon.

Table 2.6b

Percentage amino-acid composition of the centroid (see page 82).

Table 2.6a: Codon Usage of the Centroid.

UUU/Phe	118	UCU/Ser	85	UAU/Tyr	89	UGU/Cys	37
UUC/Phe	128	UCC/Ser	84	UAC/Tyr	110	UGC/Cys	48
UUA/Leu	85	UCA/Ser	72	UAA/***	..	UGA/***	..
UUG/Leu	73	UCG/Ser	38	UAG/***	..	UGG/Trp	71
CUU/Leu	83	CCU/Pro	75	CAU/His	59	CGU/Arg	88
CUC/Leu	77	CCC/Pro	63	CAC/His	72	CGC/Arg	87
CUA/Leu	67	CCA/Pro	78	CAA/Gln	98	CGA/Arg	31
CUG/Leu	185	CCG/Pro	67	CAG/Gln	138	CGG/Arg	31
AUU/Ile	156	ACU/Thr	99	AAU/Asn	111	AGU/Ser	49
AUC/Ile	148	ACC/Thr	129	AAC/Asn	156	AGC/Ser	78
AUA/Ile	60	ACA/Thr	85	AAA/Lys	189	AGA/Arg	58
AUG/Met	152	ACG/Thr	51	AAG/Lys	159	AGG/Arg	33
GUU/Val	119	GCU/Ala	167	GAU/Asp	159	GGU/Gly	159
GUC/Val	79	GCC/Ala	144	GAC/Asp	160	GGC/Gly	128
GUA/Val	83	GCA/Ala	119	GAA/Glu	209	GGA/Gly	90
GUG/Val	118	GCG/Ala	92	GAG/Glu	165	GGG/Gly	58

Table 2.6b: Amino-Acid Usage of the Centroid

Phe	4.05	Tyr	3.26
Leu	9.37	His	2.15
Ile	5.95	Gln	3.87
Met	2.49	Asn	4.37
Val	6.54	Lys	5.71
Ser	6.64	Asp	5.22
Pro	4.64	Glu	6.13
Thr	5.98	Cys	1.40
Ala	8.54	Trp	1.17
		Arg	5.39
		Gly	7.13

Opposite page 86

Table 2.6c

Percentage nucleotide composition of the centroid, broken down for each codon position.

Table 2.6c: Nucleotide Composition of the Centroid

	codon position			
	1	2	3	ALL
U/T	17.0	28.4	27.1	24.2
C	21.3	23.7	27.7	24.3
A	28.1	30.7	21.7	26.8
G	33.6	17.2	23.5	24.7
	100.0	100.0	100.0	100.0

The overall nucleotide composition of the centroid is very near to uniform usage. However the eight RNY codons account for 33.4% of the overall codon usage: this is also obvious from the nucleotide composition where R (= A & G) are over-represented in position one, and Y (=U/T & C) are over-represented in position two. The expected pyrimidine bias (PB) according to the OCAIE hypothesis (Grosjean & Fiers 1982; see discussion in chapter one) does not occur in two pairs (AU-U/C and CG-U/C) out of the eight pairs involved. There is some evidence for the expected avoidance of the seven pre-termination codons for which no synonymous alternatives exist, namely UUA, UUG, UCA, UCG, CGA, AGA, and GGA. However the latter two codons are not as infrequently used as other codons encoding the same amino-acid.

2.7.5. Conclusions.

The correspondence analysis of 428 genes detailed in this chapter has confirmed many of the findings of Grantham's group.

About one fifth of the overall variation in codon usage patterns is attributable to inter- and intra-specific variation in G+C content. There is considerable sub-structure in the cloud of 428 genes w.r.t. this factor. The first principal axis is not completely parallel with the G+C supplementary profile - this is due to the large effect that the 24 C+A rich mammalian

mitochondrial genes have in determining the position of the principal axes: a re-analysis without the eukaryotic organelle genomes resulted in a first principal axis that aligned more closely to the G+C supplementary profile (data not shown).

The pattern of G+C variation between and within species is in agreement with Sueoka (1961a): higher eukaryotes have similar mean G+C content but exhibit considerable intra-specific G+C variation; fungi and bacteria show a wide range of mean G+C content but little intra-specific G+C variation. A wide range in mean G+C content was also observed in mitochondria, in agreement with Brown (1983). Eukaryotic viruses were very similar to their hosts w.r.t. mean G+C content and intra-specific G+C variation. *E.coli* phages, however, showed no clear pattern in G+C content in relation to their host. The three large coliphage (T7, T4, and λ) differ in genomic organisation and lifestyle (i.e. lytic or lysogenic) and represent an interesting area of future study w.r.t. codon usage. After the completion of this work, Holm (1986) carried out a correspondence analysis on 78 *E.coli* genes and the pooled codon usage for the three main segments of the bacteriophage λ genome. She noted that λ genes known to be highly expressed were comparatively unbiased compared to *E.coli* highly expressed genes. She thus infers that expression level is constrained only by the rarest tRNAs. This suggests that the AGA and AGG codons are candidates for "modulator" codons.

The first principal axis is the only one of the four dimensions plotted that is easily interpretable in terms of base composition: the other three axes are dominated by three sets of genes, namely mammalian mitochondrial, *E.coli* highly expressed, and yeast highly expressed genes. The second principal axis is reflecting mainly a contrast between the characteristically biased pattern of *E.coli* highly expressed genes and the C+A rich mammalian mitochondrial genomes; the third is displaying mainly intra-specific variation in yeast; and the fourth reflects common features of *E.coli* highly expressed genes and mammalian mitochondrial genes. This last axis also appears to reflect known differences between the "universal" and the mammalian mitochondrial genetic codes.

An important factor in the codon usage pattern of mammalian mitochondrial genes appears to be high A+C content. However this is a result

of all of the genes in this study being encoded by the heavy (G+T rich) strand of the mammalian mitochondrial genome, thus resulting in A+C rich mRNAs. The one light strand gene (URF6) in the three genomes studied was not included in the analysis as it was not known to be definitely coding at the start of the study. URF6 codon usage reflects the base composition of the light strand and is correspondingly high in G+T content (data not shown). In comparison to the mammalian, yeast and *drosophila* mitochondrial genomes, the plant organelle (mitochondria and chloroplast) genes studied showed little scatter on the first four principal axes.

The positive end of the second principal axis is dominated by *E.coli* highly expressed genes. The closely related enteric bacterium *S.typhimurium* shows a similar scatter: the lack of genes with very high axis 2 scores is simply due to a low number of highly expressed genes in the *S.typhimurium* genes studied: for example, the only *S.typhimurium* ribosomal protein gene in the study, rpoD, has an axis 2 co-ordinate of 0.299. This is similar to the values obtained for *E.coli* ribosomal protein genes (a direct comparison with the *E.coli* rpoD is not possible as it was not one of the *E.coli* genes in this study). The *E.coli* plasmid and transposon genes have low axis 2 scores and show considerable variation on axis 1. This implies that they are not under the same constraints as chromosomal genes w.r.t. G+C content and are not likely to be highly expressed.

The 17 *B.subtilis* genes show some scatter on axis 2 even although it is gram-positive (*E.coli* is gram-negative) and thus very distantly related. The bias in codon usage may again to be related to level of gene expression: the *B.subtilis* genes with the highest axis 2 scores are dnaN, trpC, dnaD, rpmH and rpmA; the three lowest scores were those of spoOA, recF, and the 0.3kb. In *E.coli* and Yeast, ribosomal protein genes are very highly expressed (Ikemura & Ozeki 1983, Ikemura 1981b), and therefore B.subtilis ribosomal protein genes rpmH and rpmA would be expected to be show a similar level of expression. A recent study of codon usage in 47 B.subtilis genes (Shields & Sharp 1987) has confirmed that *B.subtilis* show a clear gradient of bias related to level of gene expression.

The four Mycoplasma capricolum genes are all ribosomal proteins and have high axis 2 scores thereby reinforcing the view that the second principal

axis reflects highly expressed genes from a wide taxonomic range of Bacteria. The other two bacterial taxons plotted, Cyanobacteria (4 genes) and Rhizobiaceae (8 genes), each consist of more than one species (two and four respectively). Little can be deduced from such limited information.

The scatter on the third principal axis is due to variation between yeast genes, and to a set of eight genes from six protozoan species. Although in these protozoan genes the inter- and intra- specific effects are confounded, the differences w.r.t. codon usage bias are worth commenting on. The three most highly biased genes are histone genes (from *T.thermophila*), the two intermediately biased genes are structural proteins, and the least biased genes are three antigen/surface protein genes. This suggests that unicellular eukaryotes may show similar intra-specific patterns of codon usage bias related to gene expression levels.

The fourth principal axis appears to reflect similarities between mammalian mitochondrial and *E.coli* highly expressed genes. Holm (1986), as noted above, has suggested that the AGA and AGG *E.coli* codons may be very susceptible to the low levels of the tRNA serving them, and may indeed be modulator codons. Thus axis 4 may be reflecting this AGA/AGG modulation factor along with the AGA/AGG mammalian mitochondrial stop codons.

RNY codons and the eight codons involved in the OCAIE hypothesis (Grosjean & Fiers 1982) only constitute part of the genetic code. It is then perhaps not surprising that their usefulness in explaining scatter on the first four principal axes is limited. However both RNY and PB factors are high at the positive end of axis 2 (*E.coli* highly expressed genes) and the negative end of axis 3 (yeast highly expressed genes). The number of codons affected by the avoidance of CpG doublets and the avoidance of pretermination codons (PTC) is eight and seven respectively. The majority of each of these sets has a G or C in the third codon position and it is therefore not surprising that these factors are related to third position G+C content.

2.8. Implications for Model Formulation.

The purpose of the correspondence analysis was to display the main features of the codon usage of the 428 genes studied and to thus to aid in

the formulation of theoretical models of codon usage.

Some of the observed trends in codon usage were already known: the differences between highly expressed and lowly expressed genes in *E.coli* and yeast; the base compositional bias in mammalian mitochondria; and the importance of G+C content. However, the analysis has shown that most of the intra-specific variation in the *E.coli* and yeast genomes is independent of G+C content, and that there are strong similarities between the codon usage patterns of the bacterial genomes studied. There is also a suggestion that other unicellular eukaryotes may resemble yeast w.r.t. codon usage.

Unicellular organisms therefore appear to have a mean G+C content with little G+C variation between genes; highly expressed genes have a more biased codon usage pattern but show little change in G+C content. Higher eukaryotes show a large amount of intra-specific G+C variation and this suggests that third position G+C content explains a large proportion of the codon usage pattern.

It is clear that there is much sub-structure in the set of genes studied. This has meant that the single correspondence analysis has focussed on many of the differences in codon usage between sets of genes. A more detailed analysis of certain sets of genes would hopefully extract aspects of the codon usage pattern that were less general. Separate analyses would therefore be more appropriate for the following subsets:

- The relationship between G+C content and codon usage in higher eukaryotes.
- The similarities in codon usage patterns between yeast and other unicellular eukaryotes.
- The similarities in codon usage patterns between *E.coli* and other bacterial species.
- The relationship between *E.coli* and the codon usage of its phages.
- The relationship between base composition and codon usage in mitochondrial genomes.
- A more extensive study of codon usage in the plant kingdom.

In addition, a general measure of the degree of codon usage bias would be useful to allow comparisons between genes. The measure used in correspondence analysis is a "chi-squared" distance from the centroid of the set of genes studied. Some consideration of what the "null hypothesis" constitutes w.r.t. codon usage would be useful.

CHAPTER 3

G+C CONTENT AND CODON USAGE IN HUMAN GENES.

3.1. Introduction.

The pioneering work of Grantham and his co-workers, (Grantham *et al.* 1980a,b), using correspondence analysis, showed third position G+C content to be the most important factor affecting the codon usage patterns of a large set of genes of different functions and from different species.

The more extensive and more detailed analysis of 428 genes carried out in the previous chapter has produced a similar result. Moreover, the intra- and inter-specific variation in third-position G+C content observed was similar to the patterns in G+C content noted by Sueoka (1961a): higher eukaryotes show little variation in the mean G+C content of their genomes, but exhibit considerable G+C variation between genes; fungi and yeast show a wide range of genomic G+C content but exhibit little G+C variation between genes. Sueoka's data refer to the G+C content of short stretches of fragmented DNA whereas Grantham *et al.* were referring to third position G+C content.

The analysis of 428 genes again appeared to confirm that third position G+C content was the main factor. A detailed examination of this factor was not carried out in the last chapter due to the considerable sub-structure in the set of 428 genes. Separate analysis of higher eukaryotes and unicellular organisms would be more appropriate.

In this chapter, a simple analysis of the relationship between codon usage and G+C content in a higher eukaryotic species (*H.sapiens*) is undertaken by considering only the G+C content in each of the three codon positions. *S.cerevisiae* is also studied to allow a comparison to be made between a unicellular eukaryotic and a higher eukaryotic organism.

3.2. G+C Content and Codon Usage.

3.2.1. Amino-Acid Usage and Synonymous Codon Usage.

Codon usage patterns can be split into amino-acid usage and synonymous codon usage. The correspondence analysis of the previous chapter revealed that 72.6% of the variation in codon usage patterns of the 428 genes studied was due to synonymous codon usage.

Analyses of codon usage patterns usually disregard the contribution of amino-acid composition. A correspondence analysis of amino-acid frequencies (Grantham *et al.* 1980a) did not reveal the clear groupings by genome type that were found by analysing codon frequencies. No significant differences in amino-acid composition are apparent between *E.coli* sequences of differing synonymous codon usage bias (Blake & Hinds 1984). However, inter-specific differences in bacteria appear to be partly due to amino-acid composition (see section 3.2.4).

Bernardi *et al.* (1985), studying G+C differences in the genes of warm-blooded vertebrates, noted that these are due to synonymous codon usage patterns alone.

3.2.2. Inter- and Intra-Specific Patterns of G+C content.

The original correspondence analyses of Grantham's group (Grantham *et al.* 1980a,b), showed third position G+C content to be the most important factor affecting codon usage patterns, although no distinction was made between inter- and intra-specific variation.

Variation in G+C content between and within organisms was reviewed by Sueoka (1961a). While Hill (1966) has noted that the overall genomic G+C content of bacteria lies between 25% and 75% (this will be denoted [25%, 75%] here), within-species variation is low (S.D. < 3%), particularly in organisms with very biased G+C content (Yamagishi 1974). Higher organisms have a much smaller range of overall genomic G+C content: higher plants [36%, 48%]; invertebrates [34%, 44%]; vertebrates [40%, 44%]. However,

intra-specific G+C variation in higher organisms, especially vertebrates, is much wider than that found in bacteria. Fungi, protozoa, and algae are similar to bacteria in their range of overall G+C content (Storck & Alexopoulos 1970). The mean and standard deviation of coding G+C content can easily be calculated from DNA sequence data: results from samples of genes from four genomes are shown in Table 3.1 (note that the observed intra-specific variation in G+C content is not due to differences in mean gene length).

Table 3.1: Intra-Specific G+C Variation in Protein Coding DNA.

Genome	No. Genes	Mean G+C %	Mean L_c	min	max
H.sapiens	135	54.1 (7.6)	297.7 (298.3)	58	2351
S.cerevisiae	110	41.8 (3.8)	331.7 (215.3)	49	1029
E.coli	50	52.7 (2.4)	302.3 (280.2)	48	1407
B.subtilis	56	44.7 (2.9)	270.8 (179.5)	43	820

Note: (1) Standard deviations are given in brackets.

(2) Gene length, L_c, is in amino-acids.

(3) *E.coli* genes taken from the correspondence analysis in chapter 2.

(4) *S.cerevisiae* and *B.subtilis* data calculated from codon usage data kindly provided by Dr. P.M. Sharp (Trinity College, Dublin).

Sueoka (1962) formulated a simple mutational model to account for the observed mean G+C content patterns noted above. The forward and back mutation rates between a G:C and an A:T pair could be altered by selection to produce an "optimal" G+C content. A similar model was proposed by Freese (1962).

The correspondence analysis of 428 genes suggests that most of the intra-specific variation in synonymous codon usage patterns in *E.coli* and

S.cerevisiae is unrelated to the G+C content of the particular genes, in complete contrast to the G+C dependent variation in higher eukaryotic genes.

3.2.3. Higher Eukaryotes.

The wide range of G+C content in vertebrate genes implies that genes of extreme base composition may have an appropriately biased amino-acid composition. Bernardi *et al.* (1985) found that amino-acid usage did not contribute to G+C differences between 34 warm-blooded vertebrate genes.

The continuing increase in available DNA sequence data has made larger analyses possible. Maruyama *et al.* (1986)'s codon usage compilation contains 135 human genes: a simple G+C analysis reveals that these genes have a G+C range of [38%, 70%]. This range is close to the theoretical limits of G+C content that can be achieved by varying synonymous codon positions only: even with extremes of G+C content in synonymous positions, the overall G+C content is constrained within the range [30.0%, 63.0%] with equal amino-acid usage, and [31.2%, 70.0%] for amino-acid usage proportional to the number of their respective codons. This latter model is more appropriate for mammalian proteins (King & Jukes 1969).

The G+C content of the third codon position of vertebrate genes is highly correlated with the "local" region of DNA in which the gene is embedded. Aota & Ikemura (1986) compared the exons, introns and flanking regions of human and chicken genes. The size of "local" region studied was between 5kb and 38kb. There is evidence that there exist much longer DNA regions of fairly constant G+C content in the vertebrate genome.

Bernardi *et al.* (1985) found that the G+C content of the coding DNA of warm-blooded vertebrates was about 10% higher than the G+C content of the local region of DNA, or "isochore",in which the gene was embedded. These isochores were very long (>>200kb), showed very little G+C variation along their length, and were of five distinct types ("components"), as classified by mean G+C content. These five components have mean G+C values of 38, 40, 43, 47, and 52 percent approximately (data taken from Bernardi *et al.* (1985): Fig 4). No precise estimates of the variation in G+C content within these isochores are available, although Bernardi *et al.* note that studies of the G+C

content of genes (i.e. the exons only) will probably fail to show significant differences between the five isochore types due to sampling error. They also note that G+C differences between isochores are less pronounced when only exons are compared.

Bernardi *et al.* concluded that this "compositional compartmentalisation" was the basis of intra-specific G+C variation in warm-blooded vertebrate genes, and that G+C differences between genes were "due to a different codon usage and not to the amino-acid composition of the corresponding proteins".

Warm-blooded vertebrate genomes have very similar mean G+C values. However, there is considerable variation in G+C content between genes. While there is some evidence that tissue-specific genes have similar G+C content (Newgard *et al.* 1986, Aota & Ikemura 1986), genes expressed specifically in different tissues are not distinguishable solely by their base composition. Newgard *et al.* (1986) have suggested that the high G+C content of muscle genes may be related to the physiological stresses observed in skeletal muscle during exercise.

The heterogeneity in DNA base composition may, alternatively, regulate expression at the transcriptional level. Bernardi *et al.* (1985) have suggested that chromosomal banding patterns are a manifestation of the isochore distribution in the genome. Thus G+C rich genes (e.g. housekeeping genes), predominantly embedded in early-replicating R bands (Goldman *et al.* (1984)), will replicate early in the cell cycle; A+T rich genes in chromosomal G bands replicate later. Early replication, while necessary, is not a sufficient condition for gene transcription. Control of expression may also be related to G+C content of the flanking regions of the gene. G+C rich flanking regions show a higher ratio of CpG to GpC than flanking regions low in G+C (Bernardi *et al.* 1985). This may affect methylation patterns and hence the control of transcription (Bird 1986).

Most authors assume that vertebrate species have similar genomes. However there may be slight differences in mean G+C content. Alonso *et al.* (1986) studied the actin gene family in the human, rat, mouse and chicken genomes. The comparison of four homologous genes between these four species suggested that there exists a gradient of G+C content from human (highest) through rodents to chicken. The difference between human and

chicken actin genes was about 5% total gene G+C (mainly due to third position differences).

3.2.4. Unicellular Organisms.

Various studies have shown that differences in G+C content between bacterial species are due, in part, to differences in amino-acid composition (Sueoka 1961b, Elton 1973). Bibb *et al.* (1984) have shown that inter-specific differences in G+C content affect all three codon positions. Although intra-specific G+C variation is low in bacteria, there is evidence of a relationship between amino-acid usage and synonymous codon usage in bacteria and bacteriophages. Wada and Suyama (1985) have shown that, within a codon, changes in G+C content in the first two positions tend to be opposed by changes at the third position.

The mean G+C content observed in bacterial species is probably under selection, via mutator genes (Cox 1976), in response to environmental pressures, e.g. U.V. light (Singer & Ames 1970), pH and temperature (Darland *et al.* 1970). Selection due to these three factors will favour high G+C content due to the higher stability of G:C pairs compared to A:T pairs. Selection for high DNA guanine-cytosine levels appears to operate on the DNA, and not on the protein produced (Ponnuswamy *et al.* 1982). A correlation between G+C content and habitat has also been found for some yeast species (Starmer & Ganter 1986), including *S.cerevisiae* which is sensitive to sunlight (Resnick 1970).

3.3. Sequence Data and Method of Analysis.

3.3.1. DNA Sequence Data used in the Study.

The codon usage data for 135 human genes was obtained from the compilation of Maruyama *et al.* (1986). An up-to-date compilation of *S.cerevisiae* codon usage (110 genes) was kindly provided by Dr. P.M. Sharp (Trinity College, Dublin).

3.3.2. Method of Analysis.

The relationship between amino-acid usage, synonymous codon usage and G+C content was studied by calculating the G+C content of each of the three codon positions. These provided rough estimates of G+C content at non-synonymous (codon positions 1 and 2), and at synonymous positions (third position).

Other estimates, correcting for leucine and arginine in the first codon position and methionine and tryptophan in the third, gave similar results. The use of G+C content measured simply over codon positions was preferred because it allowed comparisons with other studies and was amenable to simple algebraic manipulation.

Data analysis and statistical analysis programs were written in FORTRAN77 and GENSTAT respectively. All analyses were carried out on the ICL 2972 mainframe at Edinburgh University.

3.4. Results.

The overall G+C content of a gene was used as an estimate of the G+C content of the local region in which the gene lay. No direct estimate of local G+C content was available because the codon usage data was obtained from codon usage compilations (Maruyama *et al.* 1986) and not directly from sequence data.

For each of the three codon positions the regression of the G+C content of each codon position on the overall G+C content was carried out and the linear correlation coefficient calculated. The plots for the human and yeast genomes are shown in Figures 3.1 to 3.6. The slope (a) , intercept (b), and linear correlation coefficients (r) obtained for the human and yeast genomes are shown in Table 3.2.

Table 3.2: Results of the G+C Regression Analysis.

Genome	codon posn.	b	se(b)	a	r
H.sapiens (N = 135)	1	0.65	0.06	20.58	0.67
	2	0.56	0.06	10.60	0.61
	3	1.80	0.06	-31.18	0.93
S.cerevisiae (N = 110)	1	0.89	0.12	9.10	0.57
	2	0.98	0.12	-2.70	0.64
	3	1.12	0.13	-6.40	0.65

Note: (1) b, se(b), a and r represent the slope, standard error of the slope, intercept, and linear correlation coefficient, respectively, for each regression.

Due to the non-independence of individual G+C content and average G+C content, it would be expected that b is approximately equal to 1 and r approximately equal to 1/3, if the G+C content at each position varied independently and approximately equally in the three positions. The slopes are clearly significantly different from one for the human genes but are not significantly different from one for the yeast genes.

3.5. Discussion.

3.5.1. Human Nuclear Genes.

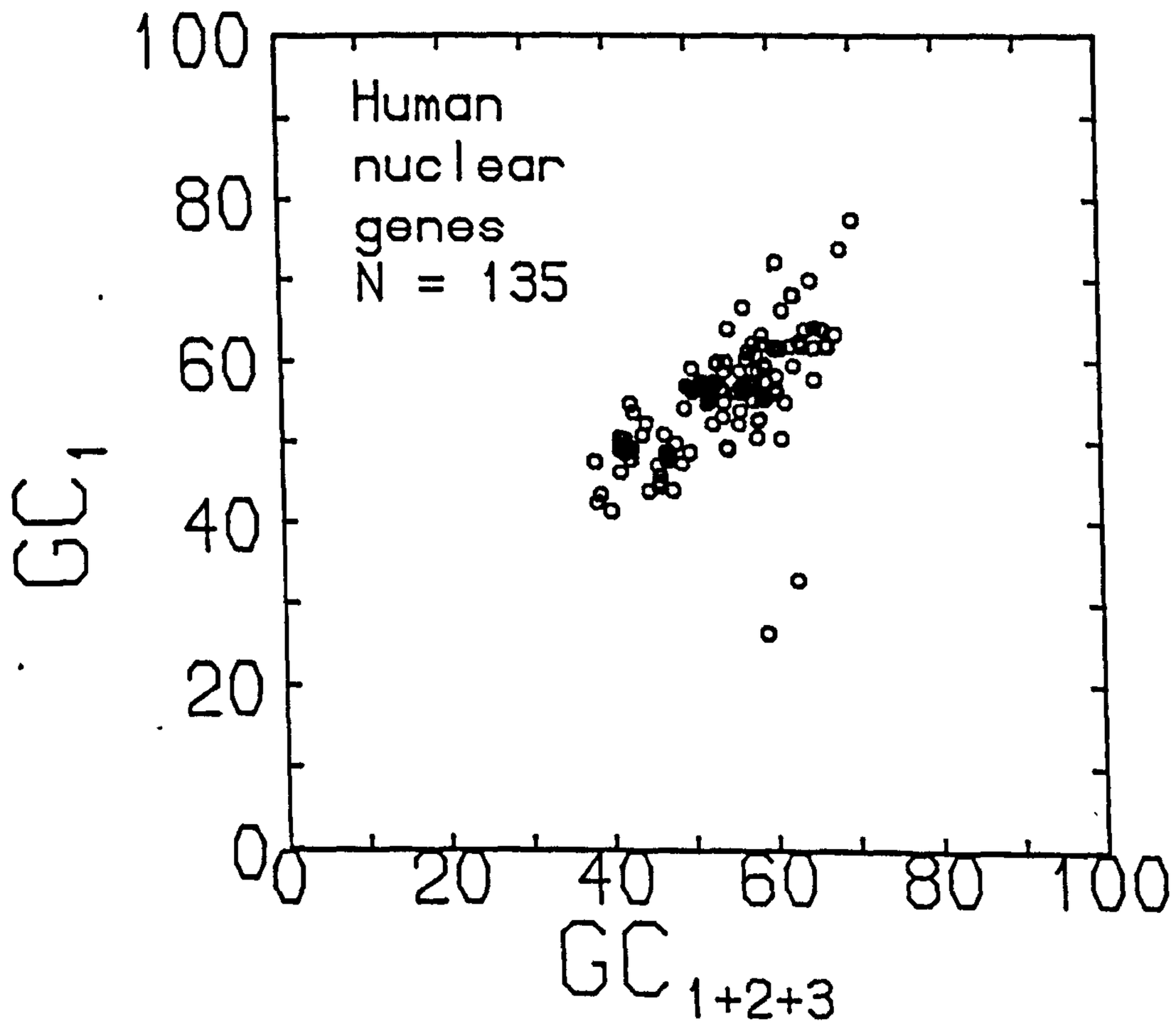
The positive correlation, in the human genome, of the first and second codon positions with total gene G+C content has important implications for molecular evolution. The first two codon positions are mainly concerned with amino-acid choice. It therefore appears that amino-acid usage is responsible for some of the G+C differences between human genes.

The finding that differences in amino-acid composition do contribute to G+C content differences between human genes is in disagreement with the results of a study of 34 genes from warm-blooded vertebrates carried out by

Opposite page 100

Figure 3.1

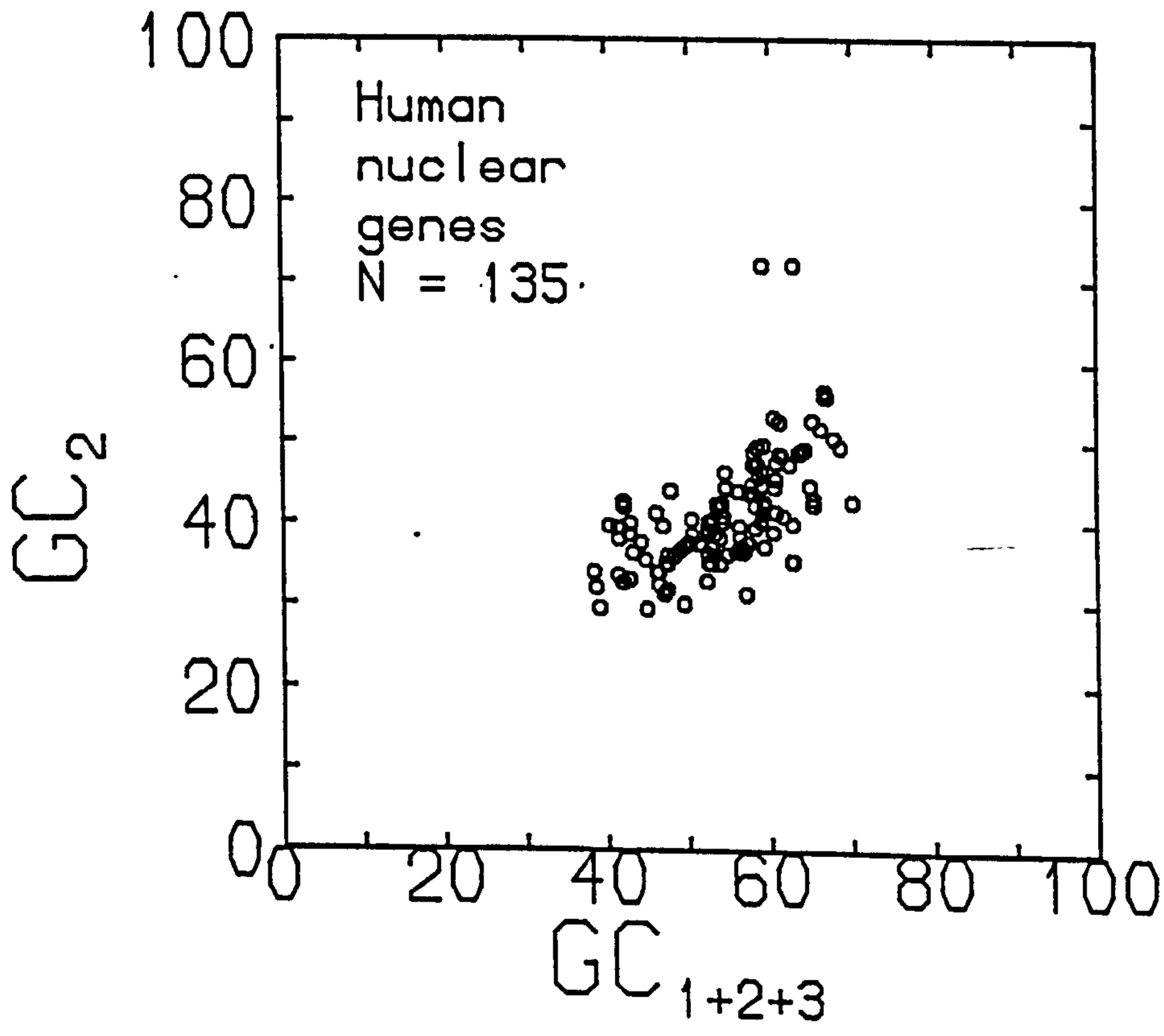
Human genes: plot of first position G+C content against the overall G+C content of the three coding positions.


Human
nuclear
genes
N = 135
100
80
60
40
20
0
0
20
40
60
80
100
GC1
GC1+2+3

Opposite page 101

Figure 3.2

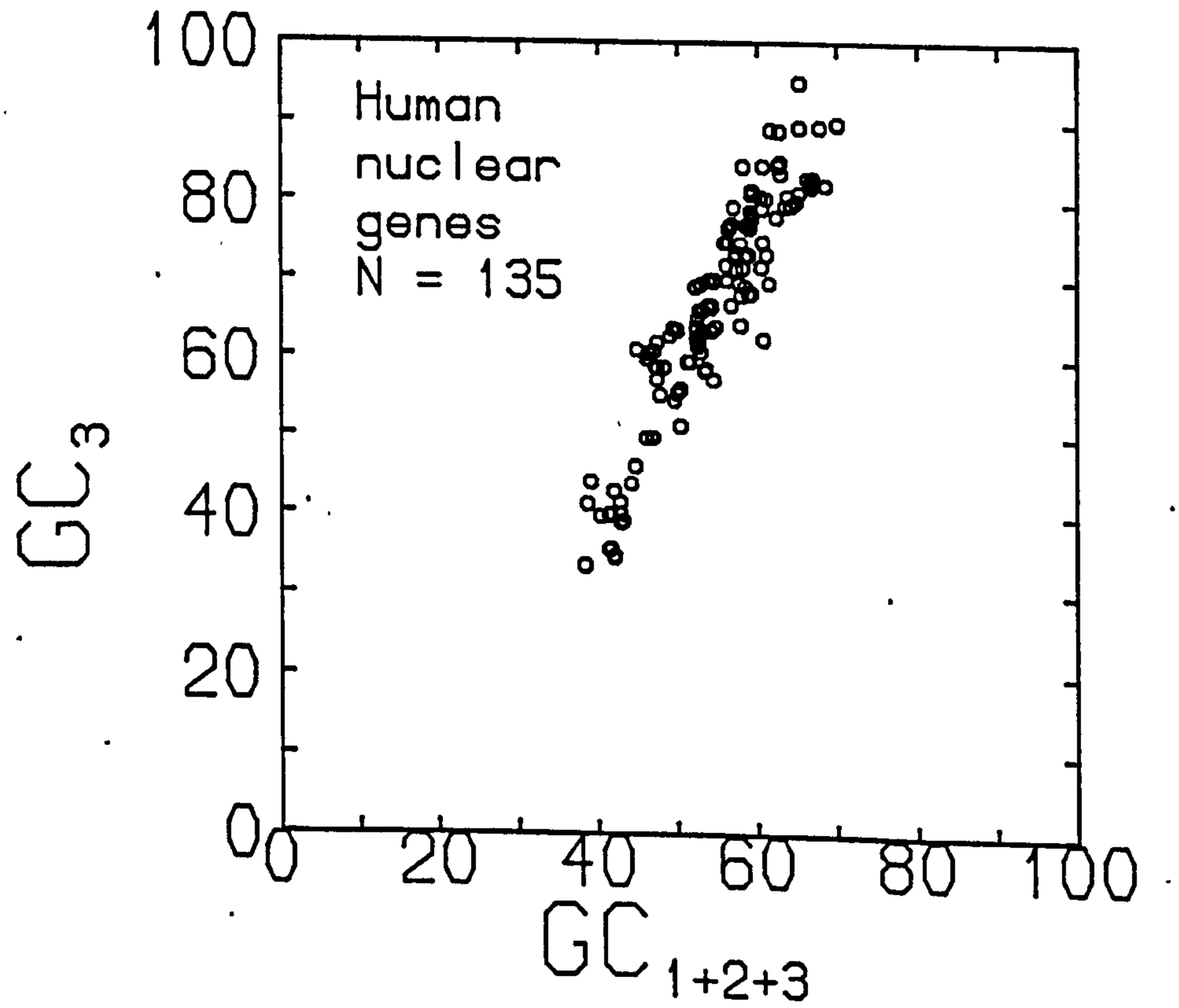
Human genes: plot of second position G+C content against the overall G+C content of the three coding positions.


Human
nuclear
genes
N = 135
GC2
GC1+2+3
0
20
40
60
80
100

Opposite page 102

Figure 3.3

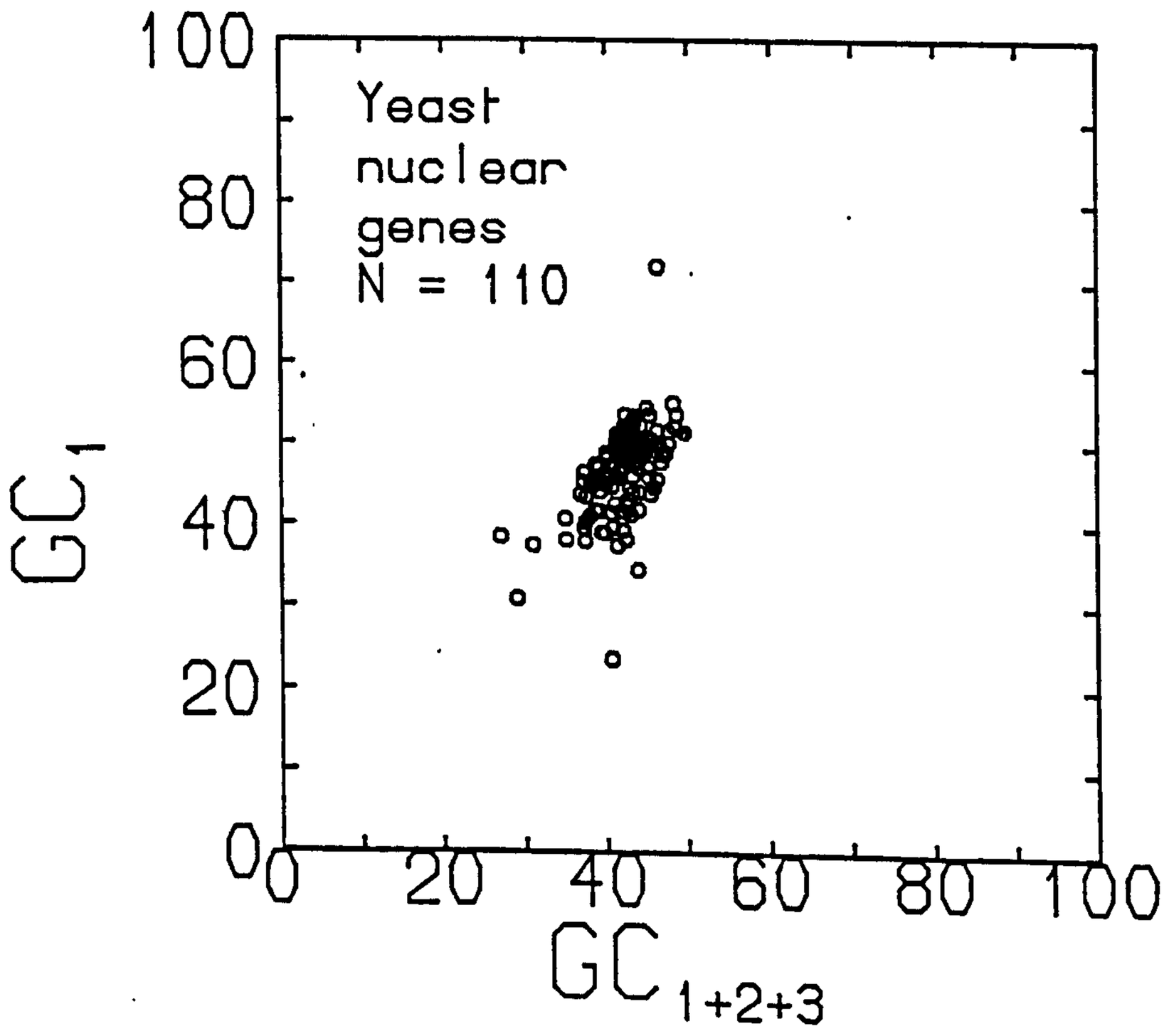
Human genes: plot of third position G+C content against the overall G+C content of the three coding positions.


Human
nuclear
genes
N = 135
100
80
60
40
20
0
0
20
40
60
80
100
GC3
GC1+2+3

Opposite page 103

Figure 3.4

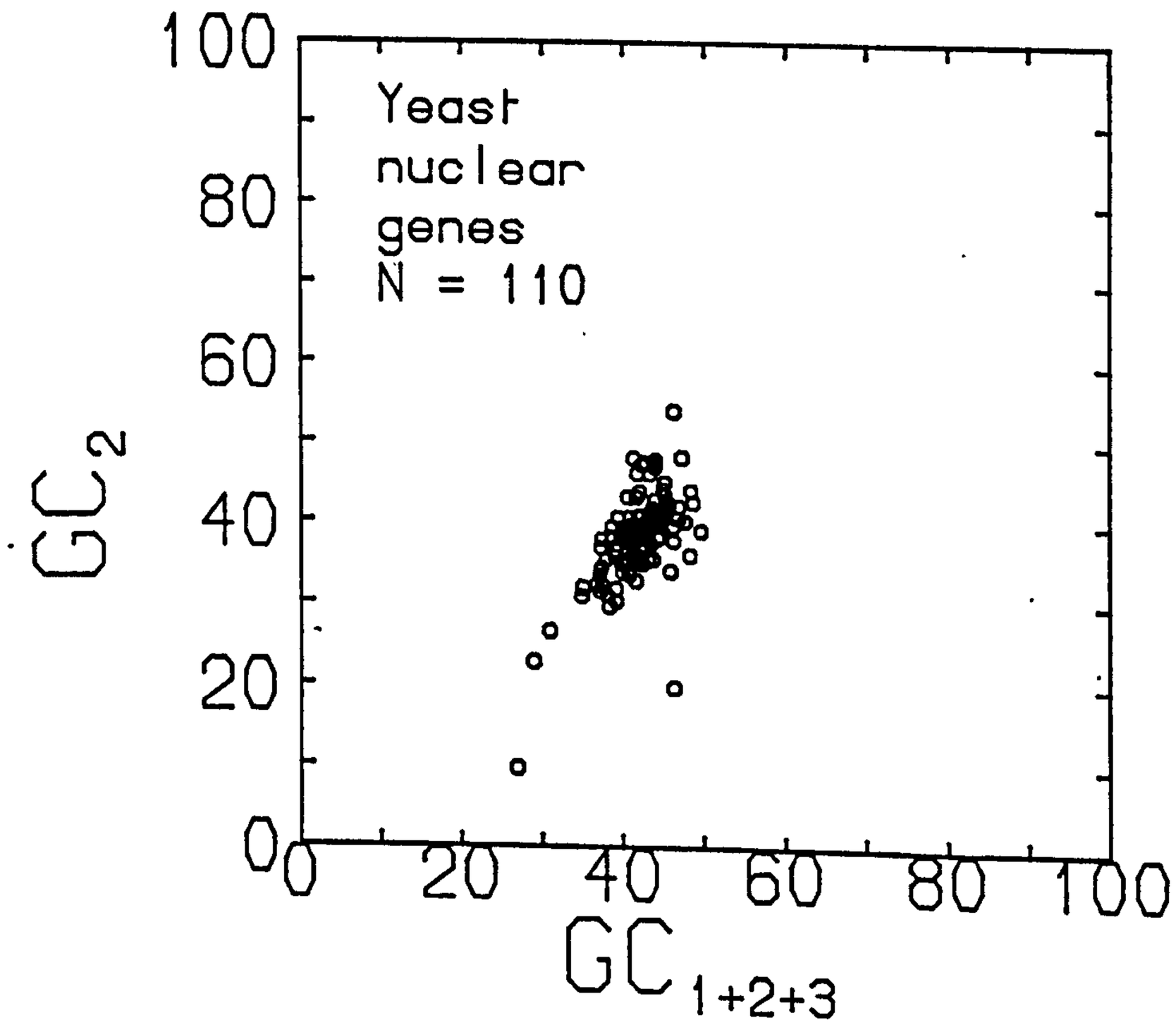
Yeast genes: plot of first position G+C content against the overall G+C content of the three coding positions.


Yeast
nuclear
genes
N = 110
100
80
60
40
20
0
0
20
40
60
80
100
GC1
GC1+2+3

Opposite page 104

Figure 3.5

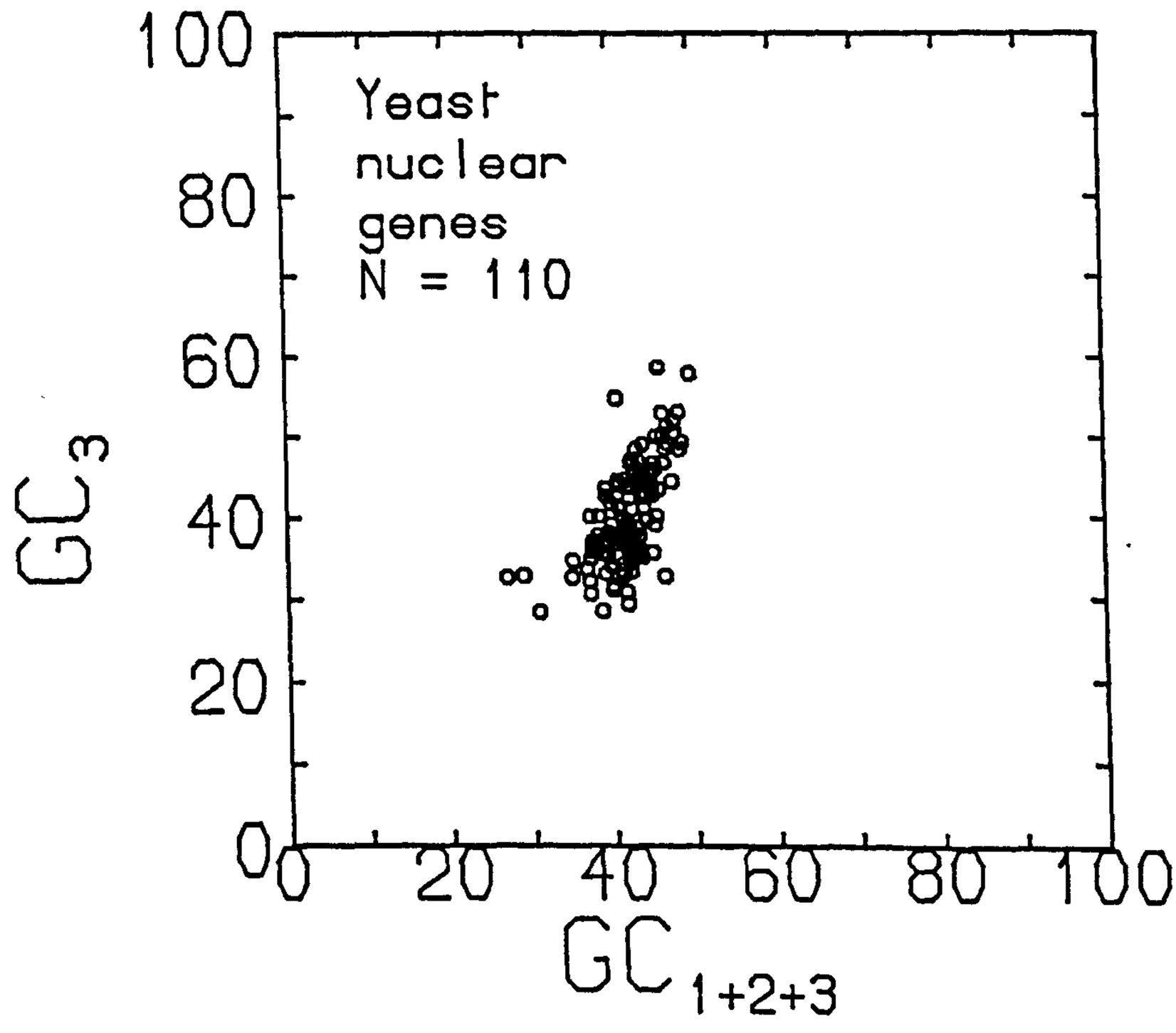
Yeast genes: plot of second position G+C content against the overall G+C content of the three coding positions.


100
80
60
40
20
0
0
20
40
60
80
100
Yeast
nuclear
genes
N = 110
GC2
GC1+2+3

Opposite page 105

Figure 3.6

Yeast genes: plot of third position G+C content against the overall G+C content of the three coding positions.


Yeast
nuclear
genes
N = 110
100
80
60
40
20
0
0
20
40
60
80
100
GC3
GC1+2+3

Bernardi · *et al.* (1985). However, they concentrated mainly on the relations between G+C content (total, and third position) and the region surrounding the gene. To emphasise their conclusion that **only** third position G+C content is involved in G+C differences between genes, they used a comparison of mouse alpha-actin genes. The authors note that cardiac and skeletal actins differ by 8% (overall) and 16% (third position) in G+C content but that there is little difference in amino-acid composition. However, simple algebra easily shows that these two genes must also differ by an average of 4% in each of the first two positions.

After the work detailed in this chapter had been completed, Bernardi's group published a subsequent paper (Bernardi & Bernardi 1986) in which they noted that compositional constraints affect **all** three codon positions. Their paper, and other recent publications, will be discussed in section 3.5.3.

To investigate the exact nature of the effect of different G+C levels on amino-acid composition, a comparison was made between a set of human genes with high G+C content and a set with low G+C content. The five genes with the highest (or lowest) G+C content formed the high (or low) G+C set. Only one member from a gene family was selected to avoid any strong influence due to gene function. The sample size of five was arbitrarily chosen. Similar results were obtained with a sample size of ten.

The amino-acid usage of the high G+C and of the low G+C sets were compared. The usage of the twenty amino-acids was calculated and expressed as a percentage of the total amino-acid composition (see Table3.3). At the bottom of the table subtotals are listed for the four columns (i.e. NUN, NCN, NAN, and NGN; where N is any nucleotide) of the genetic code table. The four columns differ only in their use of the second base in the codon. The two Serine codons in column IV of the codon usage table, i.e. AGU and AGC, were counted as belonging to column II since there are four Serine codons (i.e. UCN) in that column. This empirical decision had little influence on the results.

Opposite page 107

Table 3.3

Amino-acid composition of the "high G+C" and "low G+C" sets of human genes (see page 106 for discussion). The amino-acid composition is scaled so that equal a.a. usage would ensure a value of 1.00 for each amino acid (i.e. percentage a.a. usage divided by 5). Each amino acid is identified by its standard three letter code followed by a brief indication of which codons encode it (third position nucleotides in lower case; a '-' means all four possible nucleotides). In the lower part of the table, figures are pooled for the four columns of the codon usage table.

Table 3.3: A.A. Composition of High G+C and Low G+C Human Genes.

	High G+C (N=5)	Low G+C (N=5)	Diff.
Phe UU(u.c) :	0.53	0.84	-0.31
Leu UU(a.g) or CU- :	2.41	2.24	+0.17
Ile AU(u.c.a) :	0.37	1.22	-0.85
Met AUG :	0.46	0.64	-0.18
Val GU- :	1.14	1.30	-0.16
Ser UC- and AG(u.c) :	1.45	1.30	+0.15
Pro CC- :	1.51	0.59	+0.92
Thr AC- :	0.83	1.12	-0.29
Ala GC- :	1.93	1.02	+0.91
Tyr UA(u.c) :	0.31	0.84	-0.53
His CA(u.c) :	0.33	0.33	0
Gln CA(a.g) :	1.20	0.87	+0.33
Asn AA(u.c) :	0.33	0.87	-0.54
Lys AA(a.g) :	0.88	1.66	-0.78
Asp GA(u.c) :	0.81	1.30	-0.49
Glu GA(a.g) :	1.66	1.15	+0.51
Cys UG(u.c) :	0.44	0.46	-0.02
Trp UGG :	0.28	0.08	+0.20
Arg CG- and AG(a.g) :	1.71	1.10	+0.61
Gly GG- :	1.42	1.07	+0.35
COLUMN I (NUN) :	4.91	6.24	-1.33
COLUMN II (NCN) :	5.72	4.03	+1.69
COLUMN III (NAN) :	5.52	7.02	-1.50
COLUMN IV (NGN) :	3.85	2.71	+1.14
COLs I+II :	10.63	10.27	+0.36
COLs III+IV :	9.37	9.73	-0.36

Initial inspection of Table 3.3 reveals that within each of the four columns (i.e. NUN, NCN, NAN, and NGN), there is a tendency for the high G+C set to use amino-acids that have a G or a C in the first codon position. This is especially pronounced in the second column, where Pro and Ala are used more than (Ser + Thr). This is however not an adequate description as Ser is used slightly more in the high set. Consideration of the column totals reveals that they are not zero thus suggesting that NCN and NGN codons are used more in the high G+C set. The high G+C set (relative to the low G+C set) thus tends to use amino-acids which have high G+C content in both the first and

the second positions.

Even though Table 3.3 suggests that regions differing in G+C content tend to use amino-acids differing in the first two codon positions, this is still consistent with conservative adjustments in amino-acid usage. The first base of the codon plays a smaller role in determining the properties of the amino acid than does the second base (Alff-Steinberger 1969). The degree of functional constraint is reflected in the observed nucleotide substitution rates at the three codon positions: the third position has a higher rate than the first followed by the second (Kimura 1983). However, Sjostrom & Wold (1986) suggest that the relationships between the amino-acids are slightly more complex than can be explained by reference to the three codon positions alone. In particular, while conservative changes are likely to be produced by changes to the first codon position, this is not true for some of the column III (NAN - namely Tyr, Lys, and Asp), and column IV (NGN), codons. Also some changes to the second codon position (usually Leu <--> Pro, Val <--> Ala, and also Lys <--> Arg) do not result in drastic changes in amino-acid properties. Even with pooled data from only five genes, Table 3.3 is consistent with a model of conservative changes in amino-acid composition to buffer changes in local G+C content.

A better analysis of the amino-acid usage ramifications of G+C differences between genes would be a comparison between two members of a gene family differing considerably in G+C content. However if it can be assumed that most genes have an average amino-acid composition, some conclusions can be drawn from Table 3.3. The "conservative" differences in amino-acid composition are unlikely to be reflected in the function of the proteins produced. This is evidence that protein function is not the only factor influencing the amino-acid sequence of warm-blooded vertebrate proteins. These results suggest that G+C content as an evolutionary factor is more important than small physical and chemical differences between amino-acids.

Frommel & Holzhutter (1985) have noted that synonymous codon usage patterns can influence the rate of amino-acid substitution. This effect was investigated using the high G+C and low G+C sets of human genes. (A similar study of bacterial genes was carried out by Clarke (1983)). The only

conservative replacements likely to be affected are in A+T rich human genes. The very low third position G+C level implies very few codons ending in G. Thus amino-acid changes from UUG, CUG (Leucine), and GUG (valine), to methionine will be "pre-selected" against (Frommel & Holzhutter 1985). Therefore, methionine should be at a low level of abundance compared to G+C rich genes. However, as noted above, methionine is more abundant in genes with low G+C content suggesting that this effect does not play a large role in determining human codon usage patterns.

Bernardi *et al.* (1985)'s model of the warm-blooded vertebrate genome is one of a mosaic of DNA regions differing in mean G+C content, but with little within-region G+C variation. The heterogeneity of G+C content within a large DNA region has not been quantified. Figures 3.1 to 3.3 might have been expected to show discrete clusters of human genes if the heterogeneity was much smaller than the G+C differences between DNA regions (the 135 genes studied include a few known representatives from G+C rich and G+C poor regions). No apparent clustering is obvious.

3.5.2. Yeast Genes.

G+C differences between yeast genes are apparent in all three codon positions. Whereas most of the G+C differences between human genes are due to the third codon position, in yeast all three positions make similar contributions. This contrast appears to be due to the different proportion of synonymous codon usage bias explained by G+C content in the two species.

Only part of the total intra-specific variation in yeast synonymous codon usage is due to G+C content (see chapter 2), in contrast to human genes where G+C content appears the main factor.

This data set was analysed by Sharp *et al.* (1986). Their cluster analysis revealed two distinct groups differing in expression level and slightly in G+C content. They concluded that the observed 3 per cent difference was due mainly to G+C differences at the third position of codons.

3.5.3. G+C Content in Evolution.

The analysis of 135 human genes has suggested that both amino-acid composition and synonymous codon usage are influenced by the G+C content of the local DNA. The importance of G+C content in evolution has been stressed in a number of recent papers.

Bernardi & Bernardi (1986) have extended their model of the homeotherm genome to apply to cold-blooded vertebrates also. The authors note that both the amino-acid usage and the synonymous codon usage of a gene reflects the G+C content of the region in which it is imbedded. These results are based on an analysis of 42 vertebrate genes from 11 species, including 20 genes from the human (nuclear) genome. Their study, while dominated by human genes, does involve some inter-specific comparisons between vertebrate species. To check if consistent G+C differences exist, as suggested by Alonso *et al.* (1986) (see section 3.2.3), the slopes of Figures 3.1 to 3.3 can be compared with Bernardi & Bernardi (1986)'s results. Their values of 0.8, 0.6 and 1.6 are very similar to those obtained here (0.65, 0.56, 1.80), suggesting that vertebrates differ little in genome structure. A more comprehensive study of vertebrate codon usage is however required.

The correlation of codon usage and the G+C content of the surrounding DNA, and the possible maintenance of regions of DNA of differing mean G+C content requires a re-evaluation of our view of molecular evolution.

Bernardi & Bernardi (1986) challenge the neutral mutation random drift hypothesis (see Kimura 1983) as the main cause of evolutionary change at the molecular level, and suggest that the fixation of mutations is not at random but is "under the influence of compositional constraints . . . involving . . . both negative and positive selection".

In particular, they argue that G+C rich vertebrate genes are subject to positive selection due to the increased thermodynamic stability of their DNA, mRNA and of the protein product. This argument rests partly on the literature pertaining to the observed amino-acid replacements between mesophilic and thermophilic organisms (Argos *et al.* 1979). However, the relationship between amino-acid composition and protein thermostability is not simple and

statistical analyses of melting point data (Ponnuswamy *et al.* 1982) suggest that the object of selection is the DNA and/or mRNA, and not the protein produced.

Bernardi & Bernardi (1986) suggest that the maintenance of DNA regions of differing G+C content may be due to the action of several independent selection schemes operating on random mutational changes. This would appear to involve a very large genetic load.

Jukes & Bhushan (1986), in a study of mitochondrial and bacterial genes, also found G+C content to be an important factor. However, they proposed that the differences in G+C content between homologous genes are due to biased mutation pressure, and suggested that the forward and back mutation rates between G:C and A:T pairs are not equal. This mutational model was suggested by Sueoka (1962) to explain the observed G+C content of bacterial and fungal species (as discussed in section 3.2.2). Jukes & Bhushan (1986) suggest changes in mutational pressure are due to "shifts in the DNA polymerase system". Muto & Osawa (1987) have also used Sueoka's biased mutation pressure model to explain the G+C content of parts of the bacterial genome (tRNA, rRNA, and protein genes; spacer DNA) compared across different species. Sueoka (1986) has noted the usefulness of a "GC mutation pressure" model in explaining much of the base compositional patterns in DNA sequence data.

Extreme GC mutation pressure may help to explain the non-"universal" genetic codes found in the ciliated protozoan genus *Tetrahymena* (mean G+C content = 20% - 30%) and the bacterium *Mycoplasma capricolum* (mean G+C content = 25%). Genomes with such extremely low G+C levels may have difficulty in coding for all twenty amino acids. Jukes *et al.* (1987) propose that both *Tetrahymena* and *Mycoplasma* have evolved to use former stop codons for encoding amino-acids. For example, they suggest that *M.capricolum* may no longer use the UGA stop codon due to A+T mutation pressure, and instead use UAA. The UGA may then have been "captured", by 'wobble', as a Tryptophan codon.

Two recent papers have suggested other explanations for the G+C differences between DNA regions of varying size. Filipski (1987) has speculated that the differences in base composition between large DNA

regions (>> 200kb) are the result of the different mutational bias of alpha and beta DNA polymerases. Plants and insects only possess one polymerase, and therefore a study of codon usage in these taxonomic groups may further our understanding in this area. Adams *et al.* (1987) has suggested that periodicities in G+C content of 150–200 bp exist and that these may be related to the positioning of nucleosomes on the DNA. These comparatively short period fluctuations in G+C content, along with the (probably) relatively high heterogeneity of isochores (approximately 1.5% for mouse; Salinas *et al.* 1986) compared to the 2% to 5% differences in the mean G+C content of the five isochore types (Bernardi *et al.* 1985), make it quite unlikely that clusters of genes w.r.t. G+C content will be easily detectable.

3.6. Evolutionary Implications.

The results of this chapter and those of the recent G+C content literature require a revision of our understanding of molecular evolution and, in particular, protein evolution. The functional constraints on amino-acid positions within a vertebrate gene do not appear to be as strong as formerly suspected. However, biased mutational pressure will restrict the actual "choice" of amino-acid to a certain extent.

The possibility of different mutational pressures on different DNA regions within the warm-blooded vertebrate genome complicates our understanding of evolutionary change, and will, for example, affect estimates of genetic distance based on the assumption that the spontaneous mutation model used is applicable to the entire genome.

The observed non-uniform usage of synonymous codons was not initially expected as silent positions were thought to be neutral. Evidence has accumulated (see Ikemura (1985a,b) for reviews) pointing to selection operating on coding DNA below the level of amino-acid choice. The analysis carried out in this chapter and the work of Bernardi & Bernardi (1986) now suggests that the neutral theory applies to amino-acid choice in that small (i.e. conservative) changes in amino-acid are effectively neutral.

The relationship between codon usage and G+C content in vertebrates requires a rethink in the way that the phenomenon of synonymous codon

usage bias is viewed. The non-independence of amino-acid usage and synonymous codon usage means that a model of codon usage is more appropriate for vertebrates. A study of the relationship of amino-acid usage and other factors known to influence synonymous codon usage (e.g. factors acting on highly- and lowly-expressed *E.coli* genes) may further reveal the true nature of evolutionary forces acting on the messenger RNA.

CHAPTER 4

QUANTIFYING SYNONYMOUS CODON USAGE BIAS.

4.1. Introduction.

In chapter two, patterns of codon usage were explored using correspondence analysis. This discrete multivariate method calculates a distance, from the overall mean of the sample, for each gene. Both amino-acid composition and the usage of synonymous codons contribute to this distance. As an aid to the routine analysis of synonymous codon usage, a distance measure of the synonymous codon usage bias (SCU bias) of a gene from some stated SCU reference pattern would be useful. The aim of this chapter is to develop an estimator of SCU bias that is statistically unbiased, independent of amino-acid composition, easily computable from codon usage data only, and biologically meaningful.

4.2. Review of Measures of Codon Usage Bias.

Several measures of codon usage bias have been developed. These can be split into two main classes according to their prime objective and to the nature of the input data to the calculation. The most common class of bias measure are those designed to aid in the identification of reading frames. These measures are usually designed to operate on sequence data rather than codon usage tables and typically scan along the sequence using a window of set width. The second class of measures are those produced to quantify bias per se in known genes. These usually require only codon usage data and, in some cases, only a subset of the code is used. This classification is useful but bias measures from either class can be adapted to fulfil both roles.

Another classification is based on the choice of the synonymous codon usage pattern used as a reference. The use of the term codon usage "bias" suggests that a natural choice would be some unbiased reference SCU pattern. This unbiased pattern will be denoted as H_0. The exact details of this pattern will need to be considered further, but it can be assumed that it is easily definable. Instead of computing the distance of the SCU pattern of a given

gene from an unbiased reference, an alternative approach is to refer instead to a known biased pattern. For example, it is well-known (see Ikemura 1985a) that certain categories of *E.coli* and Yeast genes possess very biased SCU patterns. Highly-expressed genes in each of these species show very similar bias patterns, to the extent that there is a discernable overall pattern. However the two species differ considerably in this overall pattern. Therefore, if appropriate, a biased SCU reference pattern based on a priori information can be constructed. This biased pattern will be denoted by H_1. SCU bias measures based on this reference pattern can be described as one-dimensional, in the sense that genes are assumed to show a range of bias from the H_1 pattern to the H_0 pattern.

The definition of the unbiased reference SCU pattern H_0 leads to a third classification. In a study of the phenomenon of non-uniform usage of synonymous codons, a natural unbiased reference pattern is one where synonymous codons are used equally. This H_0 SCU pattern has an obvious mathematical simplicity and the notation H_0 will be retained to refer specifically to it. However, an alternative would be to set up an unbiased reference SCU pattern based on knowledge of the base composition of the gene itself, or the "local" region, or the genome as a whole. This class of reference SCU patterns will be referred to as H_0^{bc}generally.

The first quantitative analysis of codon usage patterns was carried out by Grantham *et al.* (1980a,b;1981), using correspondence analysis. This analysis was carried out on a two-way contingency table. The rows represented the mRNAs involved in the analysis; the columns, the sixty-one codons. The method does not take account of amino-acid composition and is therefore a method of analysing codon usage as opposed to synonymous codon usage. The distance metric used in this dual scaling method is proportional to the χ^2 statistic (Greenacre 1984):

$$d^2(\mathbf{p},\bar{\mathbf{p}}) = \sum(p_j - \bar{p}_j)^2/\bar{p}_j = [\ \chi^2 \text{ statistic } / \ n\ ] \qquad (4.1)$$

where $\mathbf{p}$ = vector of observed frequencies; $\bar{\mathbf{p}}$ = vector of expected frequencies. n is the absolute total for the observed vector.

This distance of the codon usage of a gene from the overall "population"

mean is therefore equivalent to a goodness-of-fit statistic over the sixty-one codons divided by n, the gene length in codons. The expected values in correspondence analysis are the mean codon usage of the set of genes in the analysis. This χ^2 distance metric weights each of the sixty-one codons by the inverse of the their occurrence in the set of genes studied so as to reduce the influence of abundant codons on the distance measure. Unweighted versions of this distance have also been discussed by Grantham *et al.* (1981) and Lipman & Wilbur (1985). Such formulations will result in the distance measure being dominated by highly used codons.

Sharp *et al.* (1986) have used a measure of SCU bias similar to the χ^2 distance from correspondence analysis. They calculate the χ^2 statistic for the twenty synonymous families independently, and then sum these tests and divide by twice the total gene length. This is essentially the statistic labelled G2 by Bishop *et al.* (1975), scaled by the total number of counts in the analysis. This distance measure is from SCU reference pattern H_0. The statistical properties of this easily computable quantity have not been investigated.

McLachlan *et al.* (1984) wished to develop a statistical test, rather than simply a distance measure of SCU bias. They discarded the possibility of using the χ^2 test due to the typically low number of counts per codon observed in real genes. This will invalidate the probabilities produced by the χ^2 test. Instead they developed a measure of SCU bias based on a multinomial standard deviation.

The distance measure developed by McLachlan *et al.* was based on taking logs of the multinomial probability of getting the actual codon usage table being studied. This quantity (multiplied by -1) was then compared to its expected value and variance calculated from generated sequences according to the desired reference pattern. The reference SCU pattern could either be H_0, or H_0^{bc} based only on base-composition in the three codon positions. Amino-acid composition was therefore not directly removed from this "SCU" reference pattern H_0^{bc}. However, in principle, any reference SCU pattern could be generated.

McLachlan *et al.*'s method has some similarity to a log-likelihood ratio approach. The "Z-score" produced by their method is:

$$[\ (-\log \hat{M}) - (-\log M)\ ] / (\text{s.d.}(-\log M)) \qquad (4.2)$$

where $\hat{M}$ is the observed multinomial probability, M is the mean (= "true") value calculated from the generated sequence data, and s.d. = standard deviation.

Their approach requires considerable computation and would be tedious without the use of a computer. However it is available as a FORTRAN77 program running on a VAX, and as part of Staden's ANALYSEQ program (Staden 1984). The program scans along the sequence using a window of 99 codons. The final "codon preference bias" statistic, CPB, is the mean value of all windows.

The first extensive comparison of a given codon usage pattern to a known pattern of bias was carried out by Ikemura (1981a). To test the relationship between tRNA abundance and codon usage, he first pooled codons known to bind to tRNAs whose abundance he had measured experimentally. The linear correlation coefficient was then calculated between this pooled codon usage (anticodon usage) and the abundance of the corresponding tRNAs. The realisation that highly expressed *E.coli* and Yeast genes had a biased SCU pattern that tended to use tRNA species which were abundant in the cell (Ikemura 1981a,b; 1982, 1985a,b), meant that a known biased SCU reference pattern H_1 could be described for these two organisms. Bennetzen & Hall (1982) devised a "codon bias index" ,CBI, based on the frequency of usage of the most preferred codons. Ikemura (1981a) developed a similar measure based on the frequency of usage of "optimal codons", f_{op}. While using a codon bias index, Bennetzen & Hall (1982) still described SCU patterns in terms of the number of codons used: e.g. highly biased yeast genes used 25 codons; relatively unbiased yeast genes used 42 codons. This informal terminology will be used in the following sections as a starting point in the development of a measure of SCU bias.

Sharp & Li (1987) have produced a SCU bias measure based on an H_1 reference pattern derived from the typical SCU pattern of highly expressed genes. This measure can be applied to any organism for which a compilation of codon usage from a set of highly-expressed genes is available, although care must be taken in using it with multicellular organisms as they appear to

have a qualitatively different type of SCU bias from *E.coli* and Yeast. The maximum absolute value of the "codon adaptation index", CAI, is obviously dependent on the degree of bias in the highly expressed reference set. The accuracy of this measure also relies on the reference set being a representative sample. The statistical properties of the CAI have not been investigated in detail although the authors report it to behave well for short genes.

Gribskov *et al.* (1984) have produced a "codon preference statistic", here denoted as CPS, that can be used with any reference SCU pattern. For each codon, the ratio of its frequency in the synonymous family is divided by the equivalent frequency in the reference SCU pattern. This preference parameter can be considered as a likelihood ratio. Unused codons are assigned low values inversely proportional to the total usage of the synonymous family to which they belong, thus preventing the ratio becoming undefined. The codon preference statistic, CPS, is then the w^{th} root of the product of these likelihood ratios, where w is the size of the window scanning the sequence. Gribskov *et al.* (1984) also consider entire genes and give examples of their method using a SCU reference set H_1 based on *E.coli* highly-expressed genes. For each of these genes they also compute the codon preference statistic using a SCU reference pattern based on an H_0^{bc} unbiased model. The range of CPS values under H_1 is from 0.59 (low bias) to 1.76 (high bias). The "control" CPS values under H_0^{bc} range from 0.44 to 0.51.

There is a final category of SCU bias measures that only describe the bias in part of the codon usage table. The P2 statistic (Gouy & Gautier 1982) is calculated only on the eight pairs of codons beginning with an A or a U in the first codon position. This statistic was developed to test the optimal codon-anticodon interaction energy hypothesis, which suggests that codon-anticodon pairs of intermediate energy are most advantageous for translational efficiency. This phenomenon is independent of tRNA abundance as each pair of codons share the same tRNA species. The P2 statistic is simply the proportion of optimal usage within each pair of codons, averaged over the eight pairs.

4.3. The Choice of a Measure of SCU Bias.

Genes vary in the degree of SCU bias. It is assumed that a particular gene of length L_c is a sample from a population of genes with a bias described by some parameter.

The intention is to develop a simple, general, easily computable measure of SCU bias that provides a conceptually obvious result. The initial idea was to produce a bias measure in terms of the number of the sense codons used, based on the terminology of Bennetzen & Hall (1982). Thus a totally biased gene would only use twenty codons (one per amino-acid). On the other hand, a totally unbiased gene would use synonymous codons equally. If the effect of amino-acid composition is removed, all the sense codons would be used equally. Such a gene could be described as using all sixty-one codons (of the "universal" genetic code) equally. Such a measure implicitly assumes an unbiased reference SCU pattern H_0.

Although the previous two chapters have provided ample evidence of the importance of base composition, and particularly G+C content, in influencing SCU patterns, the choice of such a simple SCU model has many advantages. Firstly, the measure is easily interpretable. Even when the third position G+C content differs significantly from 50 per cent, it is easy to predict the expected value of the bias measure if no other bias is present. Secondly, the mathematical simplicity of the unbiased reference SCU pattern facilitates the development of estimators of SCU bias: the analogy with Population Genetics methods for estimating homogeneity is a fruitful starting point. Thirdly, even if one wished to use an unbiased SCU reference pattern H_0^{bc}, a decision on the most appropriate value of some "base composition parameter" would be required. While such a decision is relatively easy for certain genomes, (e.g. mammalian mitochondria, human genes), in that the base composition of the codon usage table under study would be an appropriate parameter, this is not always the case. For example, the analysis of Yeast highly and lowly expressed genes would be influenced by the choice of the SCU reference pattern. The slight G+C richness of the highly expressed genes would be removed from the SCU bias pattern if the consistent third position G+C related bias was "corrected for".

The development of this measure of SCU bias will be discussed in the next section.

4.4. Development of a Measure of SCU Bias.

The phenomenon of non-uniform use of synonymous codons has some similarities with the concepts of homozygosity and heterozygosity as used in Population Genetics. In particular, the concept of the "effective number of alleles" (Kimura & Crow 1964) can be used to quantify the number of codons used by an amino-acid. The number of codons used by each amino-acid can then be combined in some appropriate way to produce a measure of SCU bias that is biologically meaningful and reflects Bennetzen & Hall's (1982) terminology.

Consider the analogy between the codon usage table and twenty different loci each with a fixed number of alleles. The codon usage table thus contains information on twenty loci (amino-acids), each possessing between one and six alleles (codons). Methods used to estimate the number of alleles in a finite population can thus be applied to the problem of quantifying the number of codons used by an amino-acid.

Kimura & Crow (1964) considered the problem of the number of alleles that could be maintained in a finite population. They used the inverse of the sums of squares of the allele frequencies to quantify the "effective number of alleles" maintained in a population. This is equivalent to the number of equally used alleles that would produce the observed homozygosity. Its value was therefore less than the actual number of alleles in the population. The effective number of alleles is the inverse of F, the proportion of homozygotes in the population.

The use of this quantity to compute the "effective number of codons" used by an amino-acid is straightforward. The maximum possible value will be the number of codons in the synonymous family. However, a decision is required on the combination of information from the twenty amino-acids. Since the main interest is in the number of codons used in the whole code, weighting each amino-acid by its number of synonymous codons is the natural choice. Two possible methods of achieving this are detailed in section

4.5.1. Both these methods produce measures of SCU bias with a theoretical range from twenty codons to sixty-one codons (for the "universal" genetic code). This class of SCU bias measures will be referred to as the **effective number of codons.**

4.5. Estimation of the Effective Number of Codons.

4.5.1. The True Value.

Two formulations of the true value of the effective number of codons are presented in this section. After studying the statistical properties of their respective estimators (see sections 4.5 and 4.6), the "best" of these two formulations will be retained.

Consider the codon usage table of a gene of length L_c codons ($3L_c$ bp). It is composed of eighteen synonymous families for each of the amino-acids which are coded for by more than one codon. Methionine and tryptophan are each coded for by one codon. The eighteen synonymous families can be split into four synonymous family (SF) types, depending on the number of synonymous codons, c_a. Note that the SF-type with c_a equal to three consists of only one amino-acid, isoleucine.

Now consider a typical synonymous family of amino-acid a consisting of c_a codons, where c_a lies between 1 and 6. The observed usage of each codon is denoted n_{ka}, and the total usage of amino-acid a, n_a, is:

$$n_a = \sum_{k=1}^{c_a} n_{ka} \qquad (4.3)$$

The observed frequency of usage of each synonymous codon, p_{ka}, is therefore:

$$p_{ka} = (n_{ka}/n_a) \qquad (4.4)$$

The most suitable statistical model is one that takes into account the categorical nature of the synonymous family. For each of the c_a categories

(codons) there is a true probability, π_{ka}. The observed n_{ka} for each of the c_a codons is then a single multinomial observation conditioned on the total n_a. An equivalent approach is to treat these observed p_{ka} as samples from a true binomial probability for each synonymous codon, π_{ka}. This is the approach adopted below in producing estimators of these π_{ka}.

Expressions are now produced for the true values of the effective number of codons for a synonymous amino-acid family N_a, and for the first overall measure of SCU bias, N_c^s. As discussed in the previous section, N_a is given by:

$$N_a = (1/\sum_{k=1}^{c_a} \pi_{ka}^2) \tag{4.5}$$

Summing these N_a for each of the twenty amino-acids yields the true value of N_c^s:

$$N_c^s = \sum_{a=1}^{20} N_a \tag{4.6}$$

The superscript s denotes that N_c^s is formed by simple summation of the N_a. The value of N_a for the single-codon amino-acids methionine and tryptophan is unity. It is convenient (see section 4.5.3) to separate their contribution to N_c^s from that of the other eighteen amino-acids. The complete expression for N_c^s is therefore:

$$N_c^s = 2 + \sum_{a=1}^{18} (1/\sum_{k=1}^{c_a} \pi_{ka}^2) \tag{4.7}$$

An alternative way of combining the $\sum \pi_{ka}^2$ from each amino-acid to produce an expression for the effective number of codons is possible. Instead of summing, over eighteen amino-acids, the reciprocal of $\sum \pi_{ka}^2$, the mean of the quantity $\sum \pi_{ka}^2$ for each of the four SF-types can be calculated, inverted, and multiplied by m_f, the number of amino-acids in the SF-type. The sum of these yields a second formulation of the true value of the (pooled) effective number of codons, N_c^p:

$$N_c^p = 2 + \sum_{f=1}^{4} (m_f^2 / (\sum_{a=1}^{m_f} \sum_{k=1}^{c_a} \pi_{ka}^2)) \qquad (4.8)$$

One of the three estimators developed ($\hat{N}_c^p$; see section 4.5.2) is designed to estimate N_c^p as defined in equation (4.8). The final decision on which formulation of the effective number of codons to use will be taken after the properties of respective estimators have been studied. When the context is obvious, the summations of p_{ka}^2 and π_{ka}^2 over k will be written in simplified form using only the summation symbol.

4.5.2. Derivation of Estimators.

In producing estimators of SCU bias, an important consideration is the finite size of coding sequences. A typical size range for mRNAs is L_c between 61 and 610 codons. Average n_{ka} values are therefore in the range 1 - 10. Asymptotically unbiased estimation is less important than the behaviour of the estimator over the above range.

Another consideration is the behaviour of estimators as amino-acid composition is varied. Biased amino-acid usage will increase sampling effects of codons in rare amino-acids especially for sparse codon usage data.

Three estimators were investigated. These were initially labelled as $\hat{N}_c^{s1}$, $\hat{N}_c^{s2}$, and $\hat{N}_c^p$. The first two seek to estimate N_c^s as defined in equation (4.7); whereas $\hat{N}_c^p$ uses the definition of N_c^p in equation (4.8). An obvious initial choice of estimator, and a useful benchmark, is obtained by replacing the true probability, π_{ka}, in equation (4.7) by the observed frequency p_{ka}. This first estimator is labelled $\hat{N}_c^{s1}$:

$$\hat{N}_c^{s1} = 2 + \sum_{a=1}^{18} (1 / \sum_{k=1}^{c_a} p_{ka}^2) \qquad (4.9)$$

Initial studies revealed that this estimator was considerably biased: as the sample size, L_c, of the codon usage table decreased, the true value of N_c^s was underestimated. This observation led to the development of two other estimators, both based on an unbiased expression for the expression:

$$\sum \hat{\pi}_{ka}^2 \tag{4.10}$$

The quantity $\sum p_{ka}^2$ can be rewritten:

$$\sum p_{ka}^2 = \sum [\, (p_{ka} - \pi_{ka})^2 + 2p_{ka}\pi_{ka} - \pi_{ka}^2 \,] \tag{4.11}$$

Taking expectations:

$$E[\, \sum p_{ka}^2 \,] = \sum E[\, (p_{ka} - \pi_{ka})^2 \,] + E[\, 2p_{ka}\pi_{ka} \,] - \pi_{ka}^2 \tag{4.12}$$

Under the assumption that n_{ka} is distributed as a Binomial (π_{ka}, n_a), the first term on the RHS of (4.12) becomes:

$$E[\, (p_{ka} - \pi_{ka})^2 \,] = \pi_{ka}(\, 1 - \pi_{ka} \,)/n_a \tag{4.13}$$

Equation (4.12) can therefore be simplified:

$$E[\, \sum p_{ka}^2 \,] = \sum [\, \pi_{ka}(\, 1 - \pi_{ka} \,)/n_a + \pi_{ka}^2 \,] \tag{4.14}$$

$$= 1/n_a + \sum \pi_{ka}^2 (\, 1 - 1/n_a)$$
$$\text{since } \sum \pi_{ka} = 1. \tag{4.15}$$

Note that:

$$1/c_a \leq \sum p_{ka}^2 \leq 1 \tag{4.16}$$

$$1/n_a \leq \sum p_{ka}^2 \leq 1 \tag{4.17}$$

Equating observation and expectation:

$$\sum p_{ka}^2 = 1/n_a + \sum \pi_{ka}^2 (1 - 1/n_a) \tag{4.18}$$

$$<=> \quad \sum \hat{\pi}^2_{ka} \;=\; (n_a \sum p^2_{ka} \;-\; 1)/(n_a \;-\; 1) \;=\; \hat{N}_a^{-1} \tag{4.19}$$

Equation (4.19) relates the estimated effective number of codons for an amino-acid, $\hat{N}_a$, to the observed synonymous frequencies, p_{ka}, and the total usage of the synonymous amino-acid family, n_a.

The sum of these $\hat{N}_a$ yields the second estimator, $\hat{N}_c^{s2}$, of N_c^s:

$$\hat{N}_c^{s2} = \sum_{a=1}^{20} (n_a \;-\; 1)/(n_a \sum p^2_{ka} \;-\; 1) \tag{4.20}$$

Since tryptophan and methionine do not possess synonymous codons, their contribution will always be unity, (see also section 4.5.3). Hence:

$$\hat{N}_c^{s2} = \; 2 \; + \sum_{a=1}^{18} (n_a \;-\; 1)/(n_a \sum p^2_{ka} \;-\; 1) \tag{4.21}$$

The third estimator seeks to estimate N_c^p as defined in equation (4.8). Substituting the $\sum \hat{\pi}^2_{ka}$ from equation (4.19) into equation (4.8) and separating off tryptophan and methionine as above, yields $\hat{N}_c^p$:

$$\hat{N}_c^p \;=\; 2 + \sum_{f=1}^{4} \left[\; (m_f^2)/\left(\sum_{a=1}^{m_f} \sum_{k=1}^{c_a} (n_a \sum p^2_{ka} \;-1)/(n_a \;-1)\right) \right] \tag{4.22}$$

$\sum \hat{\pi}^2_{ka}$ is inverted in the formula for $\hat{N}_c^{s2}$, unlike that for $\hat{N}_c^p$. This results in $\hat{N}_c^{s2}$ being undefined if the denominator of equation (4.19) goes to zero. It was this feature that led to the definition of N_c^p and the development of its estimator, $\hat{N}_c^p$.

4.5.3. Estimation when an Amino-Acid is Rare or Absent.

The formulae for $\hat{N}_c^{s1}$, $\hat{N}_c^{s2}$, and $\hat{N}_c^p$ (equations (4.9), (4.21) and (4.22) respectively) require amendment when some of the amino-acids are unused. The formulae for $\hat{N}_c^{s2}$ and $\hat{N}_c^p$ may also require adjustment when any amino-acid is used so infrequently that the usage n_a is less than or equal to

the size of the synonymous family c_a.

Missing amino-acids can be due to selection for amino-acid composition and/or the finite sample size of a gene. An estimate is required of the missing contribution of an amino-acid to the total effective number of codons. One possibility is to consider those amino-acids with the same synonymous family size, c_a, and use the average sum of the estimates of $\sum \pi_{ka}^2$ of these amino-acids as an estimate of the missing estimate of $\sum \pi_{ka}^2$. This would still require an additional decision when isoleucine is absent. This is the only amino-acid with a synonymous family size, c_a, of three. An appropriate estimate would be achieved by combining information from amino-acids with c_a = 2 and 4. In the situation where many amino-acids are absent, this method may become inoperable if all the amino-acids with a given synonymous family size c_a are absent. Besides isoleucine, this is most likely to occur with the three amino-acids which have a synonymous family size, c_a, of six.

Another possibility is to scale the information from the non-zero amino-acids appropriately. There are difficulties in deciding what proportion the missing amino-acid should make to the estimate of N_c^s or N_c^p. If a gene is totally unbiased, then the proportional contribution of a missing amino-acid would have been $c_a/61$. On the other hand, if a gene is totally biased, the same proportional contribution would be 1/20.

Empirical solutions to this problem are however available. The first estimator, $\hat{N}_c^{s1}$, can be adjusted for Z_a missing amino-acids whose synonymous codon families sum to Z_c. Note that methionine and tryptophan are always "present", and will not contribute to Z_a or Z_c. Equation (4.9) can be rewritten:

$$\hat{N}_c^{s1} = 2 + \hat{N}_c^{s1,18} \qquad (4.23)$$

where $\hat{N}_c^{s1,18}$ is the contribution to $\hat{N}_c^{s1}$ from those 18 amino-acids with synonymous families (i.e. $c_a > 1$). This formula can be generalized for use with genes with missing amino-acids by restricting the summation (detailed in equation (4.9)) to those (18 - Z_a) amino-acids with synonymous families that are present. Introducing an empirical correction for the Z_c missing codons

belonging to the missing Z_a amino-acids:

$$\hat{N}_c^{s1} = 2 + (\hat{N}_c^{s1,18-Za}) \; (59/(59 - Z_c)) \qquad (4.24)$$

The correction behaves well for synonymous families with c_a not equal to six. Detailed studies of estimator behaviour with varying amino-acid composition are carried out in section 4.6. The $\hat{N}_c^{s1}$ estimator is defined for all values of n_a, and therefore no adjustments for rare amino-acid usage are required.

The other two estimators, $\hat{N}_c^{s2}$ and $\hat{N}_c^{p}$, are influenced by rare and missing amino-acids and it is appropriate to treat these two effects together. It is clear from the equation for $\hat{N}_c^{s2}$, (equation (4.21)), that the denominator will become zero when:

$$n_a \sum p_{ka}^2 = 1 \qquad (4.25)$$

This can occur when $n_a \leq c_a$ and the non-zero codons are used once each. This includes the case where n_a is equal to one. The equation for $\hat{N}_c^{p}$, (equation (4.22)), is also undefined when any of the n_a are one. Amino-acids with n_a equal to one will be treated as missing for both these estimators.

The quantity in equation (4.25) occurs in the formulae for both $\hat{N}_c^{s2}$ and $\hat{N}_c^{p}$. However, if it takes the value one then $\hat{N}_c^{s2}$ is undefined. This event is again treated as a missing amino-acid. This decision will introduce bias into the estimate as information, possibly unbiased (in the SCU sense), is discarded. However, such events are rare and do not justify extensive investigation. $\hat{N}_c^{p}$ will also be similarly affected if all members of a SF-type are 'missing' in this way.

The treatment of missing amino-acids for $\hat{N}_c^{s2}$ is identical to the correction applied to $\hat{N}_c^{s1}$ (see equation (4.24)). The third estimator, $\hat{N}_c^{p}$, is calculated for the non-missing amino-acids. $\hat{N}_c^{p}$ is obtained (see equation (4.22)), by calculating the mean value of $\sum \hat{\pi}_{ka}^2$ for each of the four SF-types and then multiplying by m_f. There are two types of correction for missing amino-acids. If one of the amino-acids in a SF-type is missing, the mean value of the above quantity is calculated as before. If one of the SF-types is

missing or undefined (due to denominator going to zero), then the remaining contribution from the other SF-types is scaled up in a similar way to that shown in equation (4.24). Again methionine and tryptophan are always "present".

The usefulness of these empirical corrections will depend on how SCU bias is distributed between the four SF-types. The simulation design, detailed below, allows the investigation of two types of bias.

4.6. Behaviour of Estimators on Simulated Codon Usage Data.

In order to study the behaviour of the three estimators of SCU bias, a simulation program was developed. The program MULTINOM was written in FORTRAN77 and was run on the Amdahl 470 V/7 mainframe (EMAS-A) at Edinburgh University. The NAG FORTRAN library (Numerical Algorithms Group) statistical routines G05CBF and G05DAF were used for sampling random numbers from a uniform distribution.

4.6.1. Overall Simulation Design.

The program was designed to simulate a large number of codon usage tables and to calculate the bias, variance and mean square error of each of the three estimators $\hat{N}_c^{s1}$, $\hat{N}_c^{s2}$, and $\hat{N}_c^{p}$. The input to each simulation run was:

1. The amino-acid composition of the codon usage table.
2. The total number of codons, L_c, in the table.
3. The degree of SCU bias to be used.
4. The nature of the SCU bias to be used (see below).

The amino-acid composition and total number of codons were used to calculate the numbers of codons, n_a, in each amino-acid. Rounding errors were avoided in the computation of the n_a by judicial choice of the total length L_c.

The degree of SCU bias was described by the true value of N_c^s (= N_c^p).

For simplicity, this will be denoted N_c. The value of N_c ranged from 61 to 20. This was not set directly in the simulation. Instead the SCU bias was set by another parameter controlling the degree of bias within amino-acids. Desired SCU bias values were achieved by running a small iterative program, prior to the main simulation run, to compute the value of this parameter. The details of the generation of SCU bias patterns are given in the next section (4.6.2).

From this input, a "true" codon usage table was constructed of length L_c codons, with the appropriate amino-acid composition n_1 . . n_{20}, and with a synonymous codon usage pattern that would produce the required SCU bias.

From this true codon usage table, 100 samples were generated with the same total size L_c, and the same amino-acid composition. For each amino-acid the c_a multinomial probabilities, (π_{ka}, k=1,c_a), were obtained from the respective frequencies in the true table. These were then transformed to cumulative probabilities for the c_a codons, with the codons ordered as in the standard layout of the genetic code. To generate the n_a codons used by a given amino-acid a, n_a random numbers were drawn from a uniform distribution, U(0,1). These were used to allot the n_a codons to the c_a codon categories by comparing the chosen random number with the cumulative probabilities.

The three estimator formulae were then applied to these 100 sampled codon usage tables and the bias, variance and mean square error of each estimator was calculated.

4.6.2. Generation of SCU bias.

There are many ways of simulating a codon usage table with a given SCU bias and a given amino-acid composition. The contribution, to overall SCU bias, of each of the four SF-types requires consideration. If these contributions, $\sum p_{ka}^2$, do not change in a "co-ordinated" fashion as overall SCU bias changes, then the correction factors for missing amino-acids may introduce bias into the estimates. This point will be discussed in detail below.

Two methods of simulating SCU bias were used. The first again uses the analogy between twenty loci each with a finite number of alleles, and a codon

usage table. The second method was developed to produce a contrasting SCU bias pattern which distributed SCU bias among the SF-types in such a way that the absence (actual or due to low usage) of an SF-type would not result in a biased estimate of the statistic.

A consideration in both methods is how, for each synonymous family type, bias will develop. The simulation was designed so that maximum SCU bias of one codon per amino-acid was possible. While this is intuitively reasonable, it was not the only possibility. For example, extremely biased Yeast genes appear to be using a restricted set of 25 codons, in contrast to the 20 the simulation assumes. The four SF-types were parameterised as follows:

c_a	(+) codon	(-) codon(s)
2	1 + b(2)	one @ 1 - b(2)
3	1 + b(3)	two @ 1 - b(3)/2
4	1 + b(4)	three @ 1 - b(4)/3
6	1 + b(6)	five @ 1 - b(5)/4

The relationships between the b(k) are detailed below.

Crow & Kimura (1970) provide a relationship between the effective number of alleles, n_e, and the quantity $M = 4N\mu$ for loci with k alleles. Applying this to synonymous families, the effective number of codons per amino-acid for each of the four non-unity synonymous family types is:

c_a	N_a
2	(2M + 1)/(M + 1)
3	((3/2)M + 1)/((1/2)M + 1)
4	((4/3)M + 1)/((1/3)M + 1)
6	((6/5)M + 1)/((1/5)M + 1)

The first method of generating SCU bias uses these relationships to define the

relative bias between synonymous families. This method will be denoted the "M" method. The relationships between the bias parameters, b(k), were appropriately constrained. The actual degree of SCU bias is then set by the quantity M.

The second method constrains the b(k) so that:

$$c_a \sum \pi_{ka}^2 = \text{constant}$$

Under this constraint, the maximum bias cannot reach $N_c = 20$. The maximum value of the above expression is two for the case where $c_a = 2$. This constrains the larger SF-types e.g. for $c_a = 6$, the most extreme bias possible will be an $N_a = 3$. The minimum value of N_c is 31.5 codons.

The mean value of the above quantity was calculated for each of the four SF-types, on the simulated data and for a compilation of *E.coli* codon usage data. The results were displayed as a plot of $c_a \sum \pi_{ka}^2$ versus the synonymous family size c_a. Method W will produce lines with a gradient of zero; method M will produce a positive gradient. The *E.coli* data used (Sharp & Li 1986) consists of compilations of very highly-expressed (34.1), highly-expressed (40.3), regulatory (50.2), and other (46.7) genes. The $\hat{N}_c$ values are given in brackets.

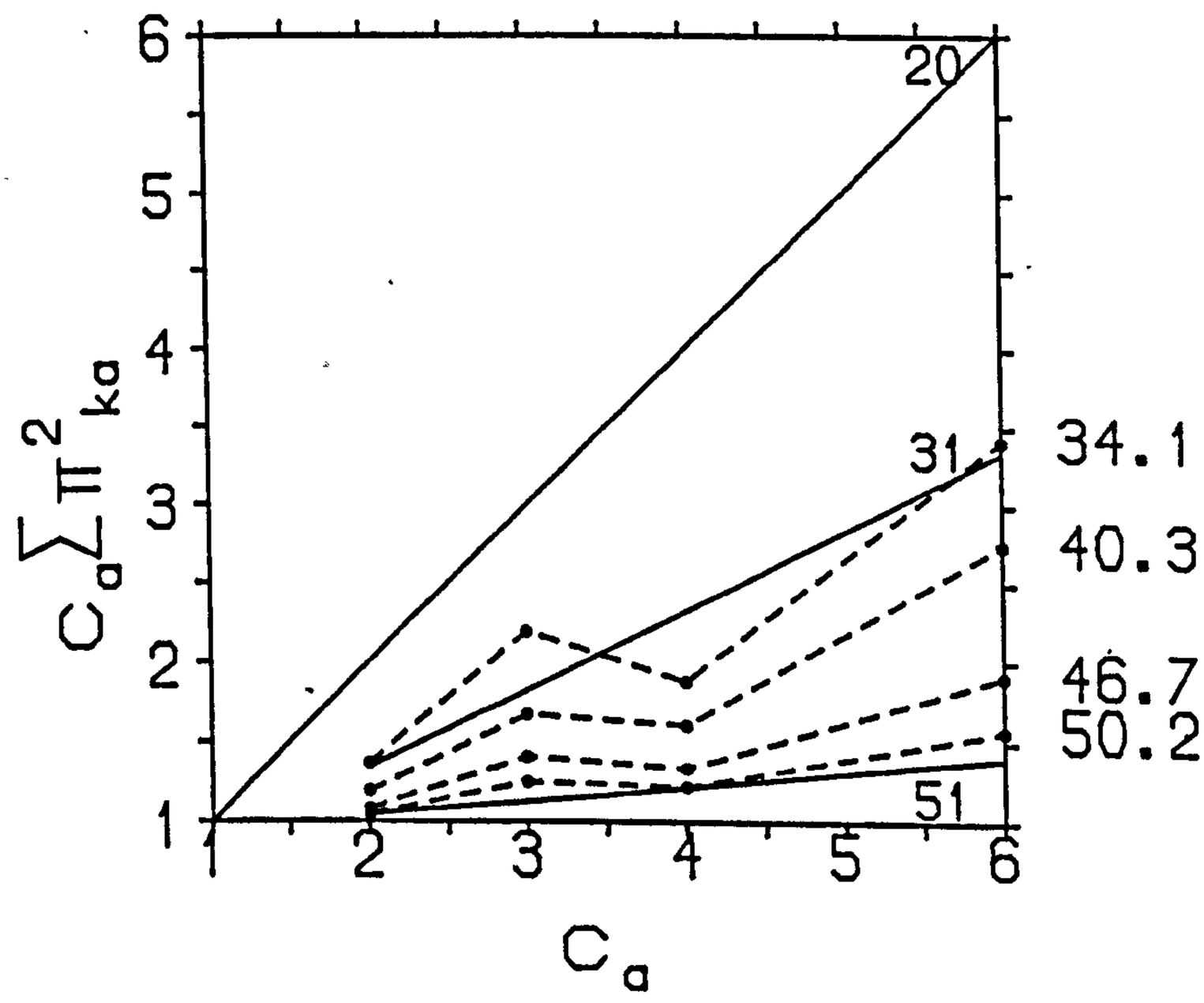
This information is shown in figure 4.01. The line with slope unity represents total bias ($N_c = 20$, as noted at top right of figure). No bias would be represented by the X-axis. The two other solid lines are produced by simulating bias (for $N_c = 31$ and 51) using method M. The four dashed lines represent the four *E.coli* compilations, identifiable by their $\hat{N}_c$ values on the right of the figure.

It is clear that SCU bias simulation method M is a reasonable approximation to the distribution of SCU bias within categories of *E.coli* genes. Similar results were obtained for Yeast highly and lowly expressed genes. A slope of zero is consistent with SCU bias due to G+C content: extreme bias of this type will lead to only $c_a/2$ codons being used per amino-acid (except for isoleucine, and the two amino acids with $c_a = 1$). This is similar to method W. These considerations confirm that real genes are well described by both the methods of simulating SCU bias.

Opposite page 132

Figure 4.01

Generation of SCU bias: comparison of pooled *E.coli* codon usage data (dotted lines) and SCU bias generated by methods W and M (see page 131). The figures on the far right of the figure describe the SCU bias of the *E.coli* data. See discussion on page 131.


6
5
4
3
2
1
2
3
4
5
6
20
31
51
34.1
40.3
46.7
50.2
$C_a \sum \Pi^2_{ka}$
C_a

4.6.3. Simulation Results.

The simulation analysis was carried out using methods W and M to generate SCU bias. Two amino-acid compositions were used:

ECU: amino-acid composition proportional to the synonymous family size, c_a, of each amino-acid. This is the composition that equal codon usage would produce, hence ECU.

EAAU: equal amino-acid usage.

The size of the codon usage table studied, L_c, was chosen to reflect the size of real coding sequences. Thus the range used for ECU was 61 to 1220; for EAAU it was 60 to 1200. The incremental size was chosen so that no rounding errors would occur in the calculation of amino-acid composition. Nine L_c values were investigated.

For the W and M methods, the following ranges of SCU bias (in terms of N_c) were generated:

Method W: 61 51 41 31.5

Method M: 61 51 41 31 21

The behaviour of the three estimators, under two methods of simulating SCU bias and two amino-acid compositions, is displayed graphically in figures 4.02 to 4.13. The estimator under study is identified on the Y-axis. The true values of N_c are shown on the right of each plot; the SCU bias simulation method and the amino-acid composition are stated at the top. The error bars represent the square root of the variance of each estimator for a given value of L_c. L_c is the size of the codon usage table in codons.

Initial inspection of the simulation results reveals that estimator $\hat{N}_c^{s1}$ behaves very poorly (see figures 4.02 - 4.05). The underestimation of N_c^s is considerable for moderately and lowly biased genes, ($N_c^s \geq 41$), over a range of L_c typical of real coding regions, ($L_c < 600$ codons). This estimator is thus discarded. The rest of this discussion will concentrate on the two other

estimators. The behaviour of $\hat{N}_c^{s2}$ is shown in figures 4.06 - 4.09; $\hat{N}_c^{p}$ in figures 4.10 - 4.13.

Figure labelling:

	M		W	
	ECU	EAAU	ECU	EAAU
$\hat{N}_c^{s1}$	4.02	4.03	4.04	4.05
$\hat{N}_c^{s2}$	4.06	4.07	4.08	4.09
$\hat{N}_c^{p}$	4.10	4.11	4.12	4.13

A measure of the performance of an estimator is the mean square error, MSE. For example, the MSE of the estimator $\hat{N}_c^{s2}$ w.r.t the true value N_c^{s} is:

$$\mathrm{MSE}(\hat{N}_c^{s2}) = E[\ (\hat{N}_c^{s2} - N_c^{s})^2\ ] \tag{4.26}$$

$$= V(\hat{N}_c^{s2}) + [b(\hat{N}_c^{s2})]^2 \tag{4.27}$$

where b($\hat{N}_c^{s2}$) is the bias of $\hat{N}_c^{s2}$.

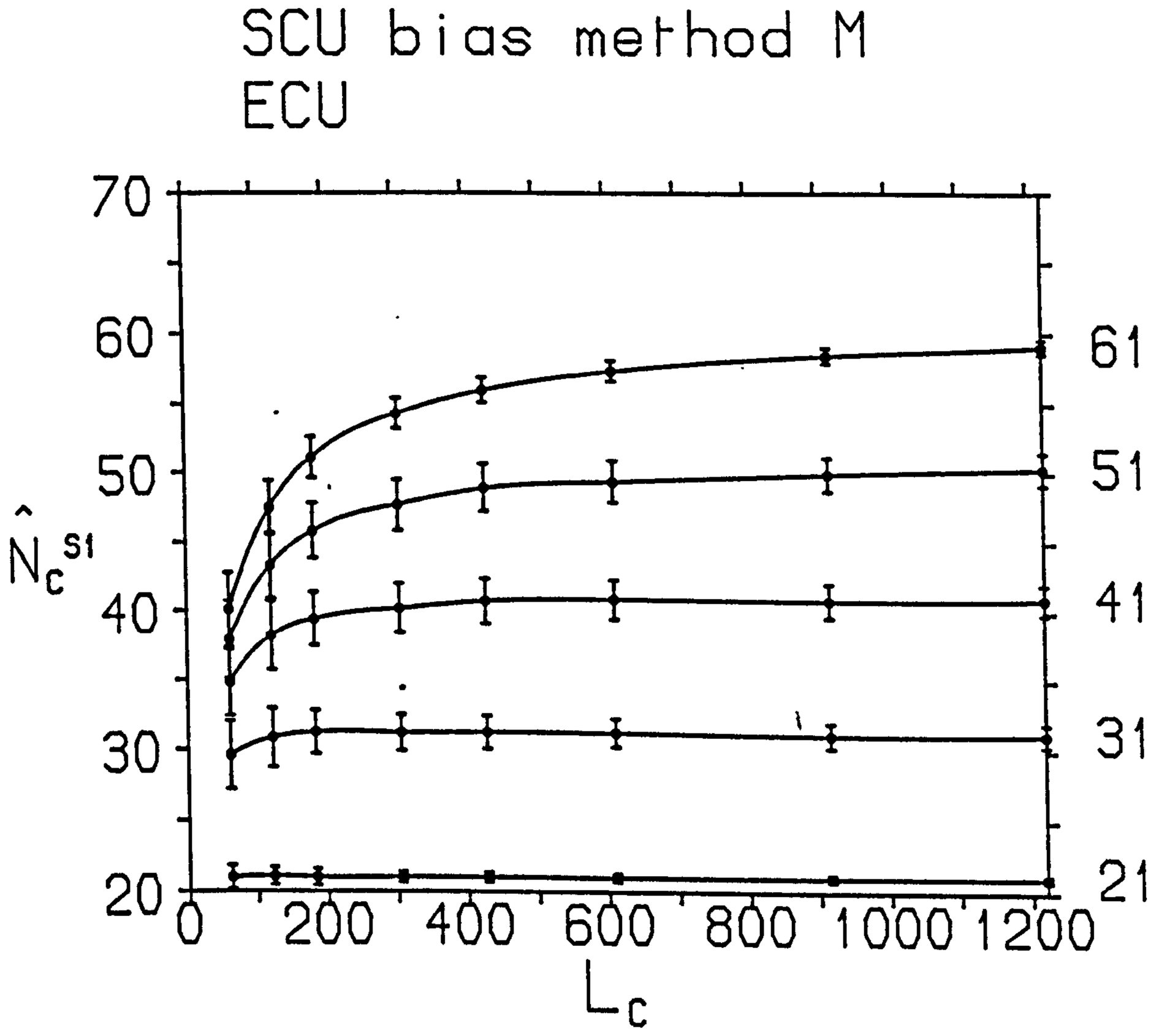
The relative performance of $\hat{N}_c^{p}$ and $\hat{N}_c^{s2}$ is similar in the four simulations (M/ECU, M/EAAU, W/ECU, AND W/EAAU). For the two estimators, there are only small differences in their respective variances for a given value of L_c and a given amino-acid composition. Table 4.2 shows the variance ratios (relative efficiency) of these two estimators for the M/ECU simulation. Most of the differences in the MSEs of $\hat{N}_c^{p}$ and $\hat{N}_c^{s2}$ are due to differences in bias (e.g. see M/ECU bias data in table 4.3). Similar variance ratios are obtained for the M/ECU, W/ECU and W/EAAU simulations. The effect of different amino-acid compositions leads to an increase in the respective MSE values. For $L_c > 180$ codons, this is due mainly to an increase in variance. For lower L_c values, an increase in bias is obvious when the EAAU results are compared to those for ECU (see the MSE data in tables 4.1 (M/ECU) and 4.4 (M/EAAU), and figures 4.06 to 4.13).

Both estimators behave poorly at L_c equal to 60 or 61 codons. The MSE value for $\hat{N}_c^{p}$ is very high due to the condition outlined in equation (4.25)

Opposite page 135

Figure 4.02

Estimator behaviour as a function of gene length, L_c. The method of generating SCU bias (i.e. either W or M), the type of amino acid composition (i.e. either ECU or EAAU), and the estimator under study are all detailed in the figure. see pages 133 and 134 for further details.

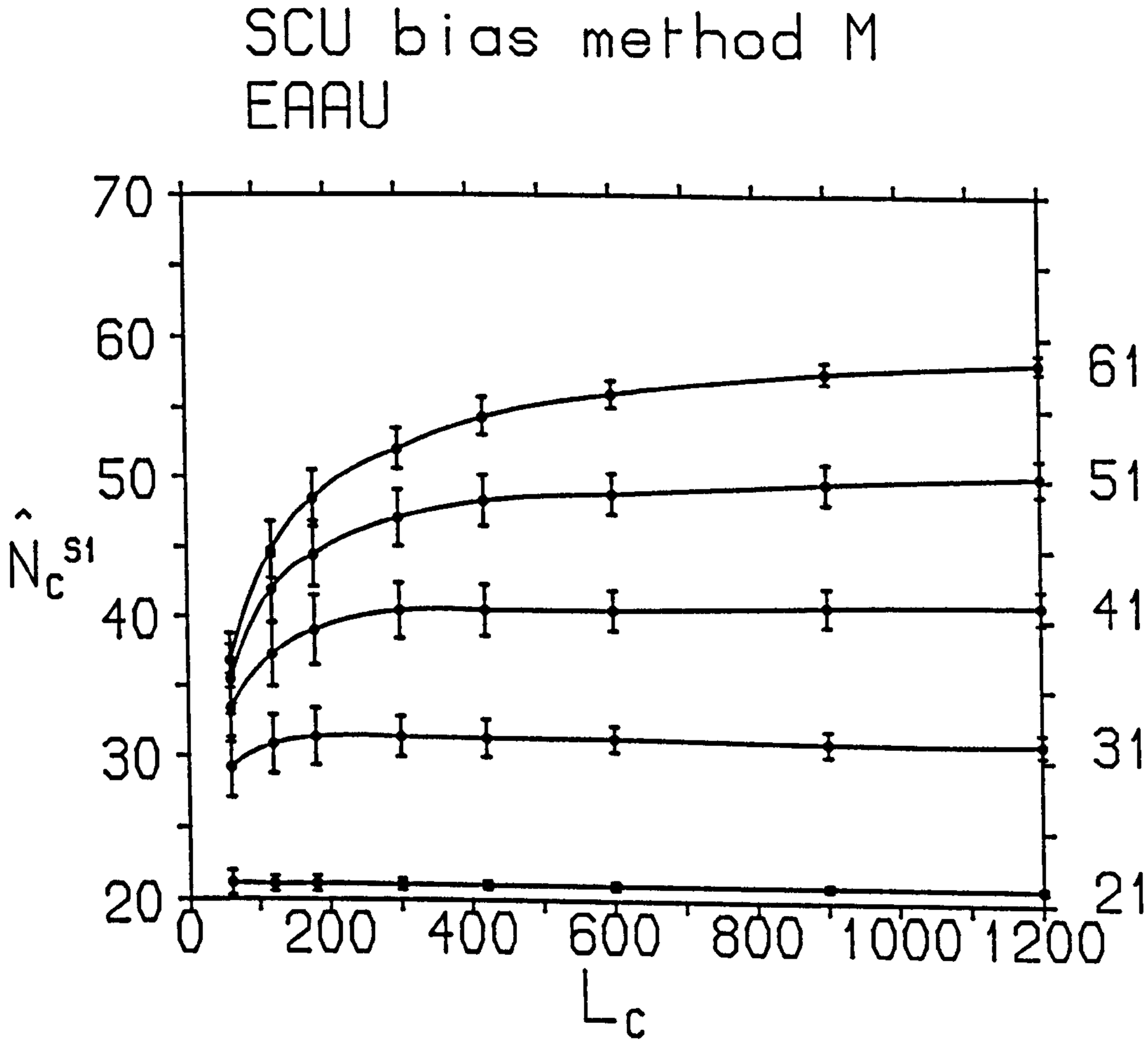

SCU bias method M
ECU
70
60
50
40
30
20
$\hat{N}_c^{s1}$
61
51
41
31
21
0
200
400
600
800
1000
1200
L_c

L

Opposite page 136

Figure 4.03

Estimator behaviour as a function of gene length, L_c. The method of generating SCU bias (i.e. either W or M), the type of amino acid composition (i.e. either ECU or EAAU), and the estimator under study are all detailed in the figure. see pages 133 and 134 for further details.

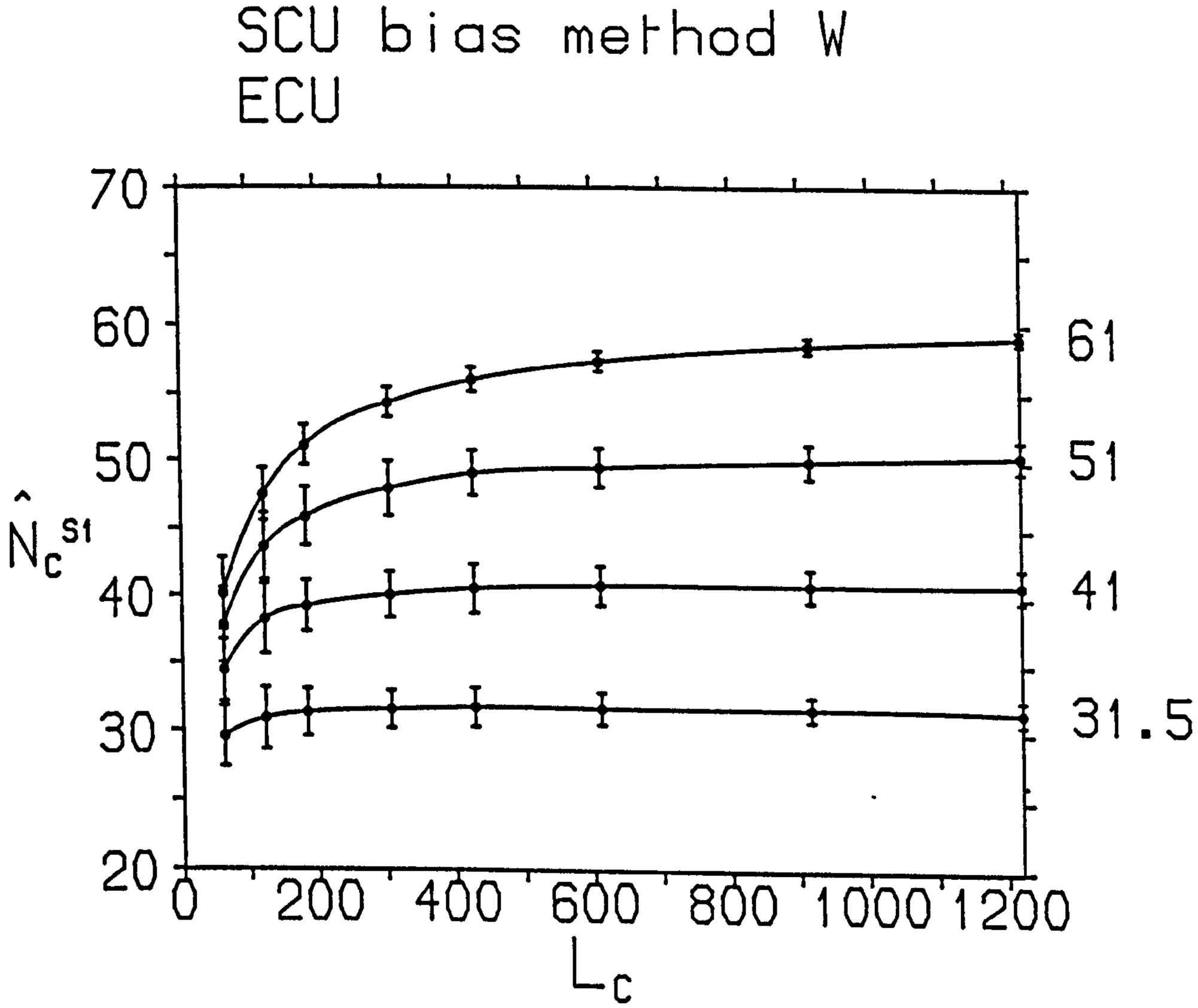

SCU bias method M
EAAU
70
60
50
40
30
20
$\hat{N}_c^{S1}$
0
200
400
600
800
1000
1200
L_c
61
51
41
31
21

L

Opposite page 137

Figure 4.04

Estimator behaviour as a function of gene length, L_c. The method of generating SCU bias (i.e. either W or M), the type of amino acid composition (i.e. either ECU or EAAU), and the estimator under study are all detailed in the figure. see pages 133 and 134 for further details.

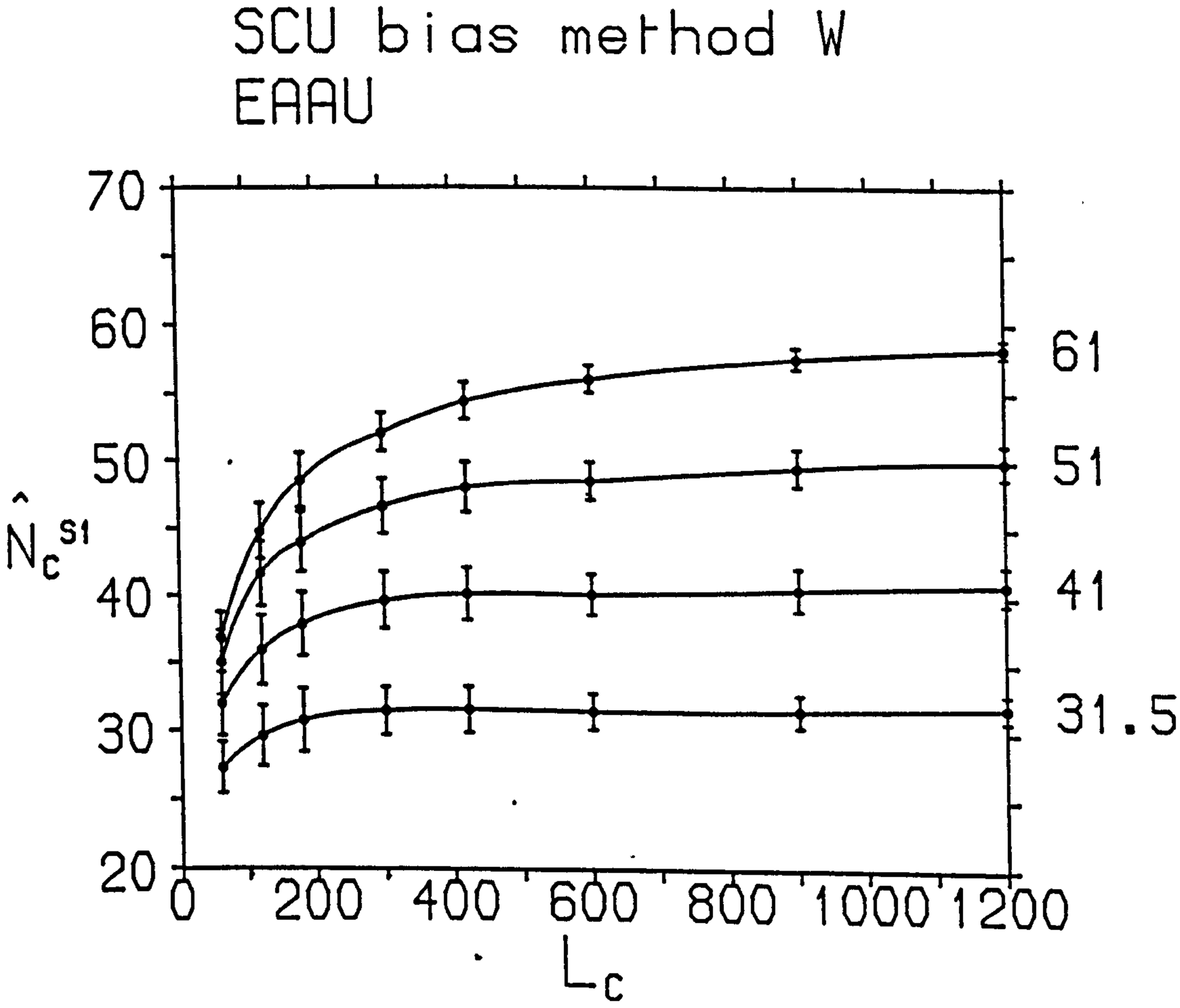

SCU bias method W
ECU
70
60
50
40
30
20
$\hat{N}_c^{S1}$
61
51
41
31.5
0
200
400
600
800
1000
1200
L_c

L

Opposite page 138

Figure 4.05

Estimator behaviour as a function of gene length, L_c. The method of generating SCU bias (i.e. either W or M), the type of amino acid composition (i.e. either ECU or EAAU), and the estimator under study are all detailed in the figure. see pages 133 and 134 for further details.

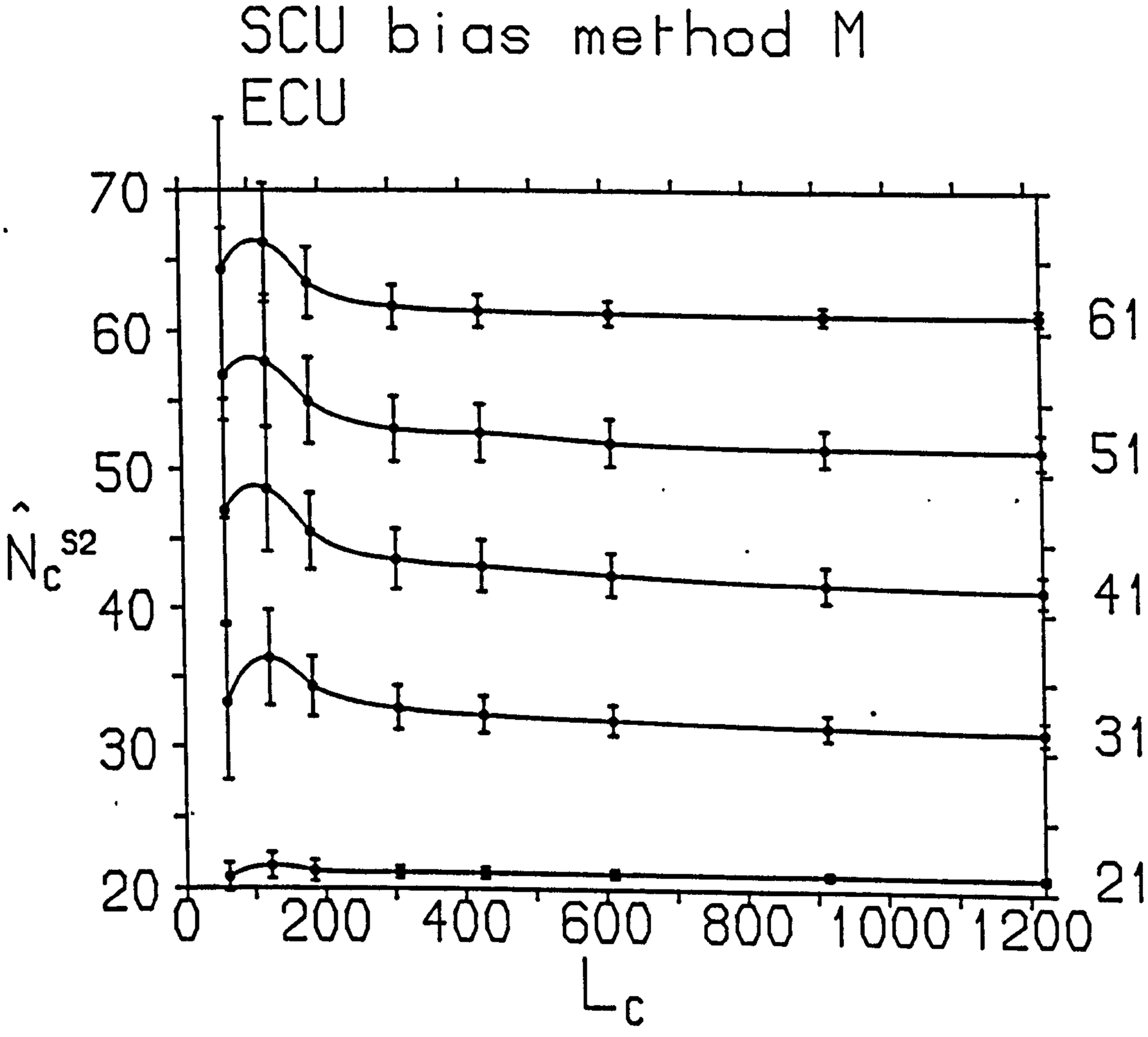

SCU bias method W
EAAU
70
60
50
40
30
20
$\hat{N}_c^{S1}$
61
51
41
31.5
0
200
400
600
800
1000
1200
L_c

L

Opposite page 139

Figure 4.06

Estimator behaviour as a function of gene length, L_c. The method of generating SCU bias (I.e. either W or M), the type of amino acid composition (i.e. either ECU or EAAU), and the estimator under study are all detailed in the figure. see pages 133 and 134 for further details.

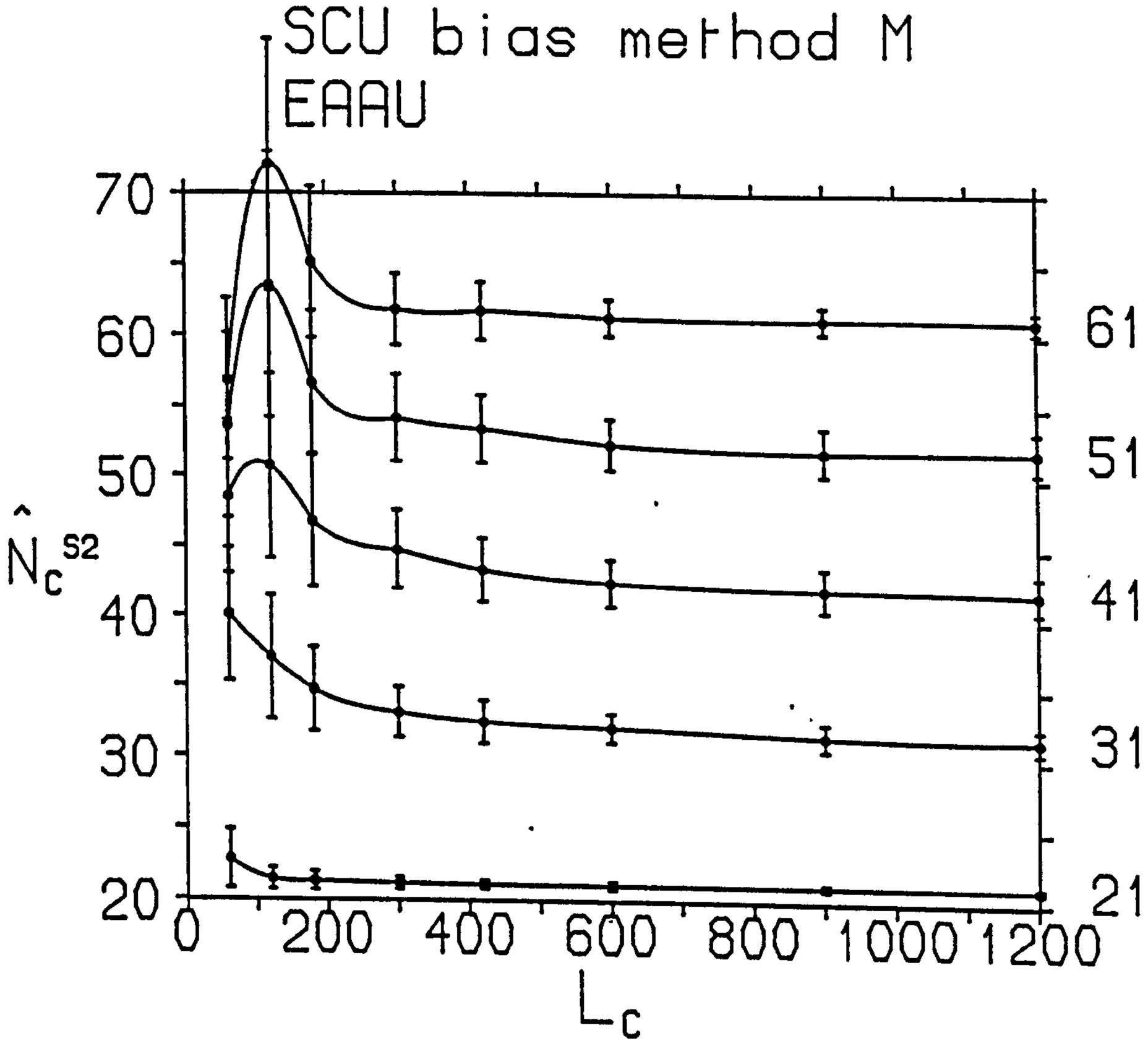

SCU bias method M
ECU
70
60
50
40
30
20
$\hat{N}_c^{S2}$
61
51
41
31
21
0
200
400
600
800
1000
1200
L_c

L

Opposite page 140

Figure 4.07

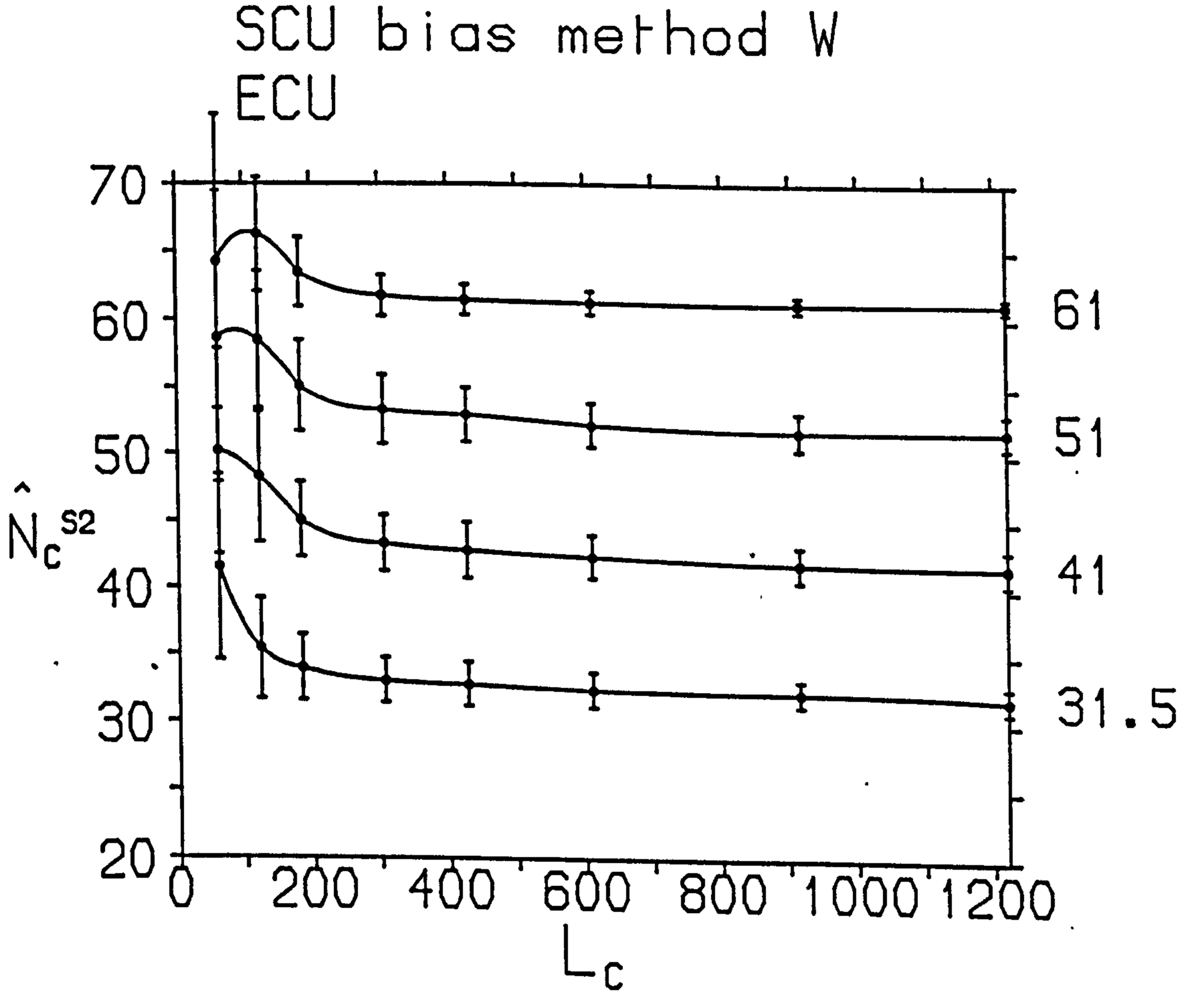
Estimator behaviour as a function of gene length, L_c. The method of generating SCU bias (i.e. either W or M), the type of amino acid composition (i.e. either ECU or EAAU), and the estimator under study are all detailed in the figure. see pages 133 and 134 for further details.


SCU bias method M
EAAU
70
60
50
40
30
20
$\hat{N}_c^{S2}$
0
200
400
600
800
1000
1200
L_c
61
51
41
31
21

Opposite page 141

Figure 4.08

Estimator behaviour as a function of gene length, L_c. The method of generating SCU bias (i.e. either W or M), the type of amino acid composition (i.e. either ECU or EAAU), and the estimator under study are all detailed in the figure. see pages 133 and 134 for further details.

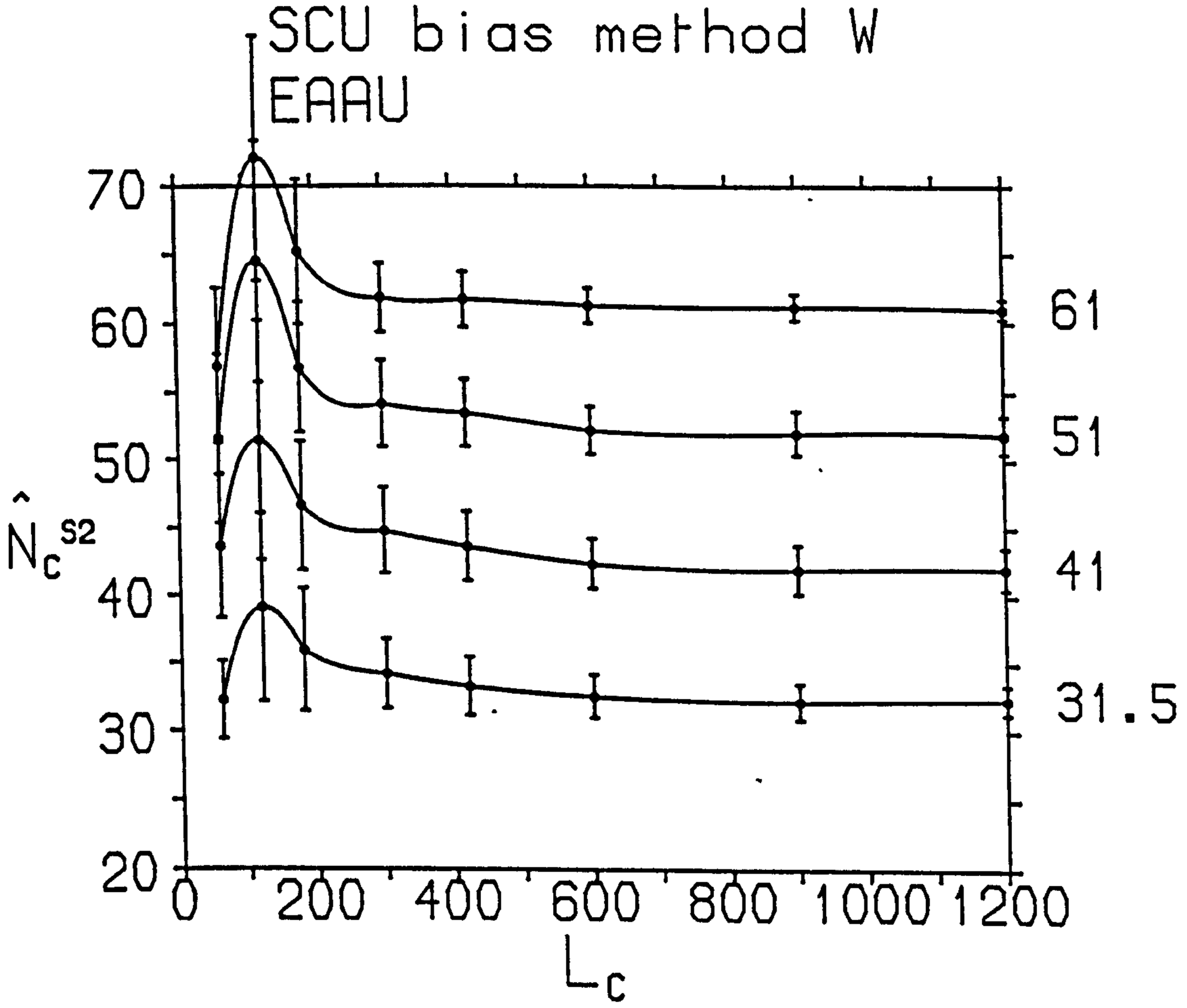

SCU bias method W
ECU
70
60
50
40
30
20
$\hat{N}_c^{S2}$
0
200
400
600
800
1000
1200
L_c
61
51
41
31.5

L

Opposite page 142

Figure 4.09

Estimator behaviour as a function of gene length, L_c. The method of generating SCU bias (i.e. either W or M), the type of amino acid composition (i.e. either ECU or EAAU), and the estimator under study are all detailed in the figure. see pages 133 and 134 for further details.

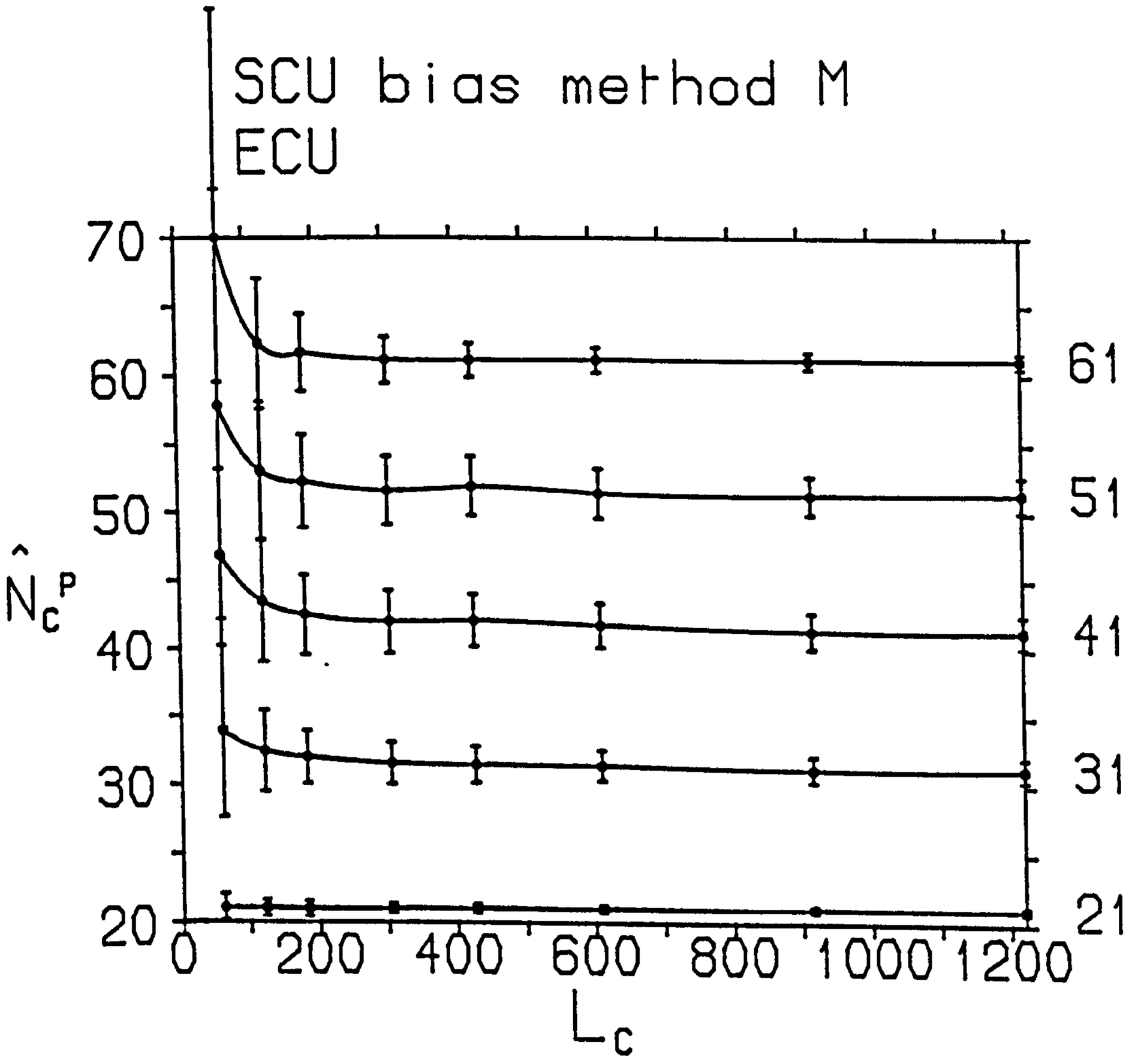

SCU bias method W
EAAU
70
60
50
40
30
20
$\hat{N}_c^{S2}$
0
200
400
600
800
1000
1200
L_c
61
51
41
31.5

L

Opposite page 143

Figure 4.10

Estimator behaviour as a function of gene length, L_c. The method of generating SCU bias (i.e. either W or M), the type of amino acid composition (i.e. either ECU or EAAU), and the estimator under study are all detailed in the figure. see pages 133 and 134 for further details.

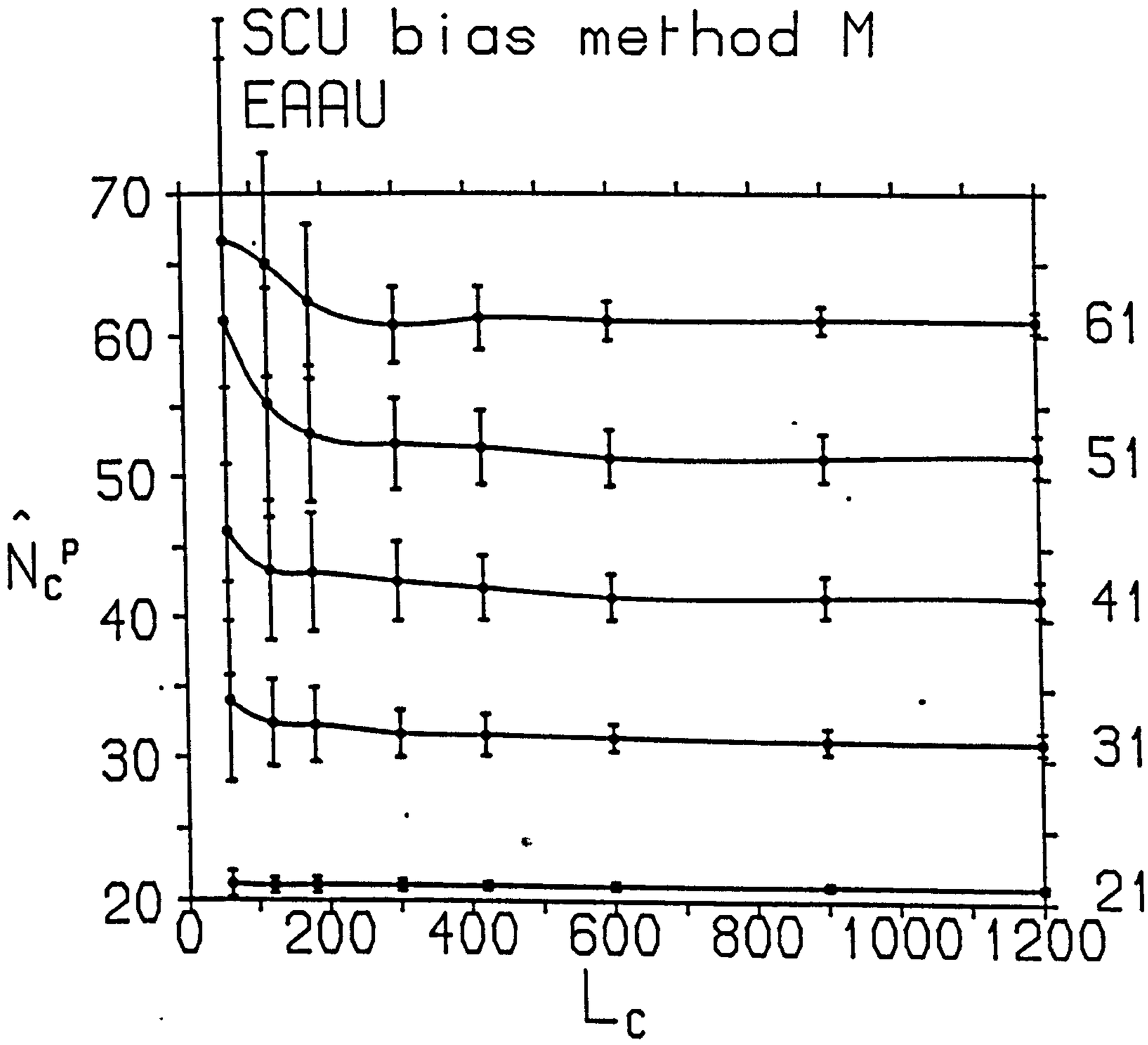

SCU bias method M
ECU
70
60
50
40
30
20
$\hat{N}_c^p$
61
51
41
31
21
0
200
400
600
800
1000
1200
L_c

L

Opposite page 144

Figure 4.11

Estimator behaviour as a function of gene length, L_c. The method of generating SCU bias (i.e. either W or M), the type of amino acid composition (i.e. either ECU or EAAU), and the estimator under study are all detailed in the figure. see pages 133 and 134 for further details.

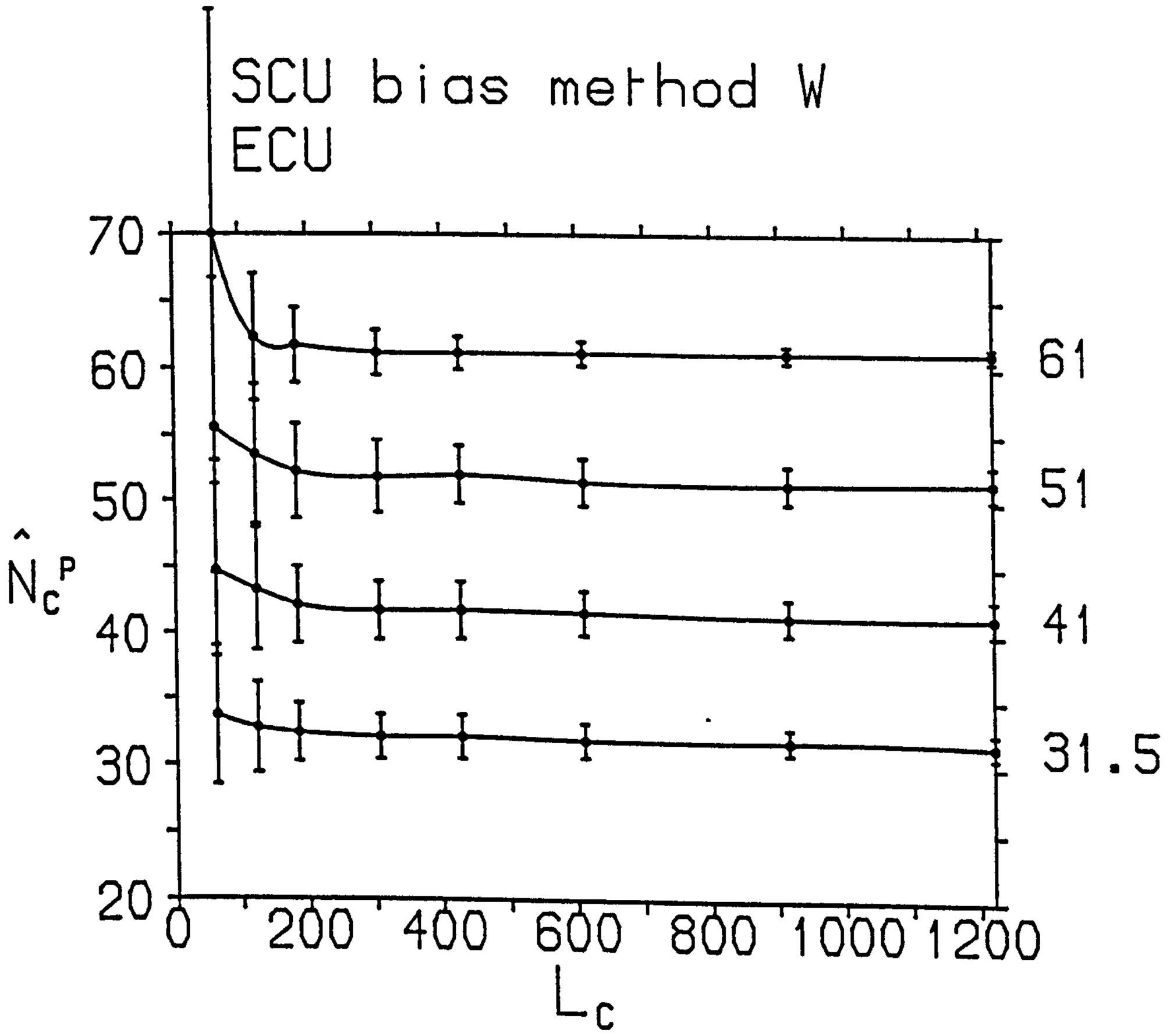

SCU bias method M
EAAU
70
60
50
40
30
20
$\hat{N}_c^P$
61
51
41
31
21
0
200
400
600
800
1000
1200
L_c

L

Opposite page 145

Figure 4.12

Estimator behaviour as a function of gene length, L_c. The method of generating SCU bias (i.e. either W or M), the type of amino acid composition (i.e. either ECU or EAAU), and the estimator under study are all detailed in the figure. see pages 133 and 134 for further details.

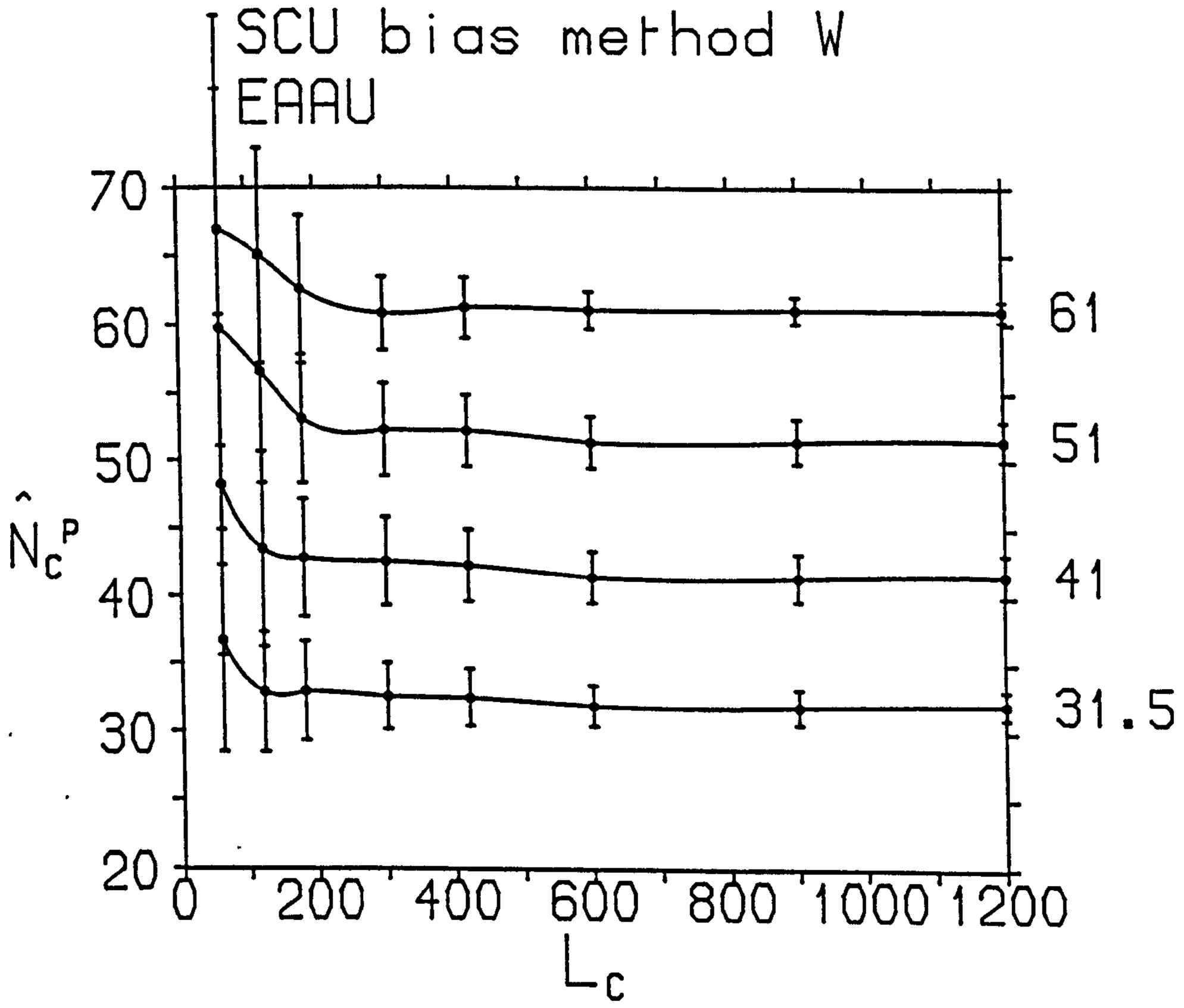

SCU bias method W
ECU
70
60
50
40
30
20
0
200
400
600
800
1000
1200
$\hat{N}_c^P$
L_c
61
51
41
31.5

L

Opposite page 146

Figure 4.13

Estimator behaviour as a function of gene length, L_c. The method of generating SCU bias (i.e. either W or M), the type of amino acid composition (i.e. either ECU or EAAU), and the estimator under study are all detailed in the figure. see pages 133 and 134 for further details.


SCU bias method W
EAAU
70
60
50
40
30
20
0
200
400
600
800
1000
1200
61
51
41
31.5
$\hat{N}_c^P$
L_c

L

holding for several amino-acids. The equivalent values for $\hat{N}_c^{s2}$ are lower due to excluding these amino-acids.

The two methods of simulating SCU bias (methods M and W) had very little influence on the values of the estimators. Even with very small codon usage tables (L_c = 60/61), there was small differences. A comparison of $\hat{N}_c^{s2}$ and $\hat{N}_c^{p}$ is shown in figures 4.14 and 4.15.

$\hat{N}_c^{p}$ is clearly a better estimator of SCU bias than $\hat{N}_c^{s2}$. Neither estimator behaves well on data from very short genes, but $\hat{N}_c^{p}$ shows low bias over the gene lengths typical of real genes.

4.7. Application to Real Codon Usage Data.

The $\hat{N}_c^{p}$ estimator was calculated on a representative sample of fifty *E.coli* and fifty *S.cerevisiae* genes. The codon usage of these two species has been the subject of considerable study (see Ikemura (1985b) for a review). Most of the previous work on the development of measures of SCU bias have studied genes from these two species. This facilitates comparison with the estimator developed in this chapter.

E.coli Codon Usage.

The *E.coli* results are shown in table 4.5 and have been sorted according to the value of $\hat{N}_c^{p}$. Other information supplied is the length L_c (codons), overall G+C content, the χ^2 statistic divided by L_c, and the class of gene. This latter classification is:

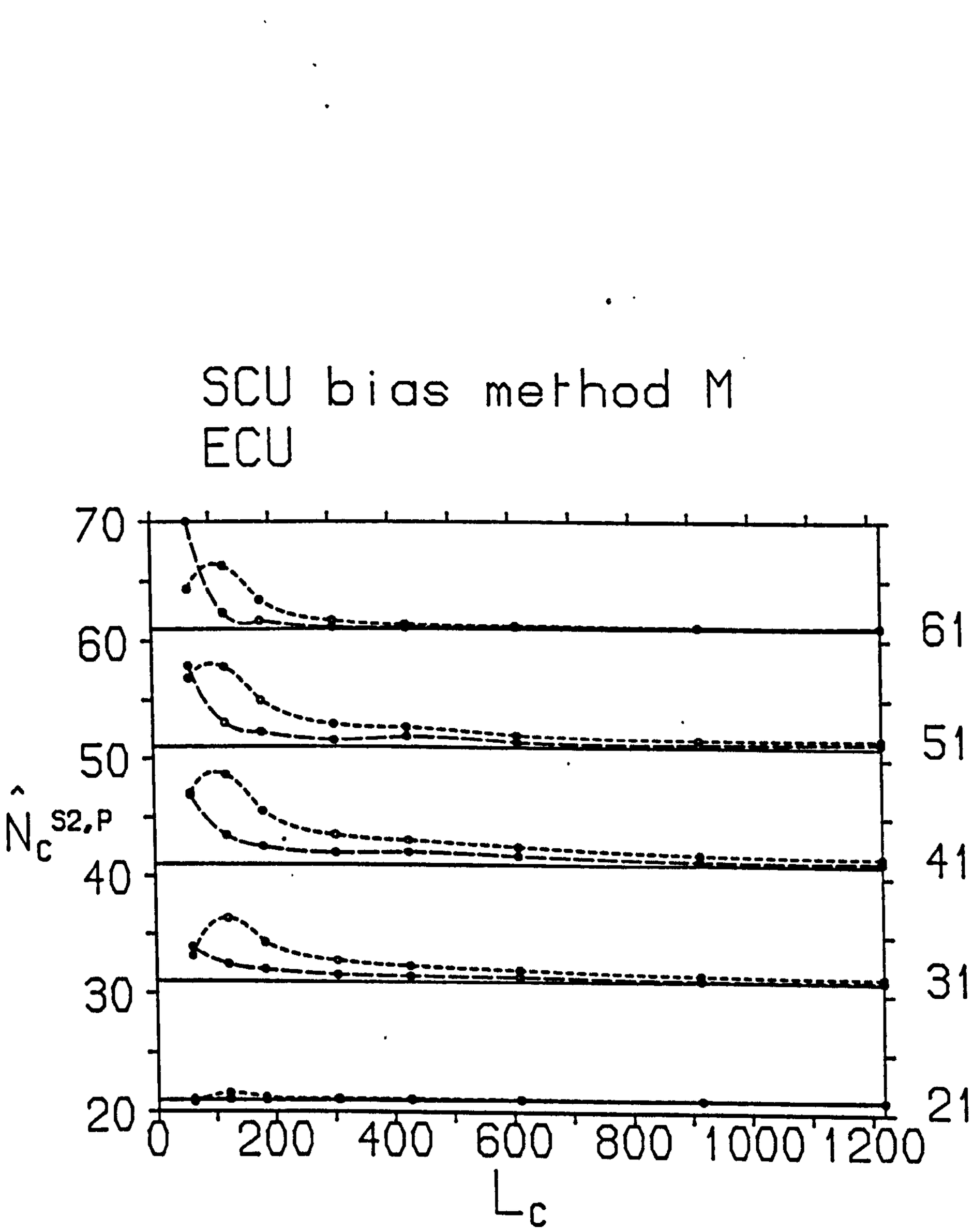
```
 VH  very highly expressed
  H  highly expressed
MOD  moderately expressed
REG  regulatory gene
 TP  transposon or plasmid gene
```

Information on the expression level of *E.coli* genes was obtained from Sharp & Li (1986). Additional information on the values of four other measures of bias is provided as discussed in section 4.2.

Opposite page 148

Figure 4.14

Comparison of estimators. $\hat{N}_c^p$ is identified as those lines with a longer dash. See pages 133–134 and 147 for further details.

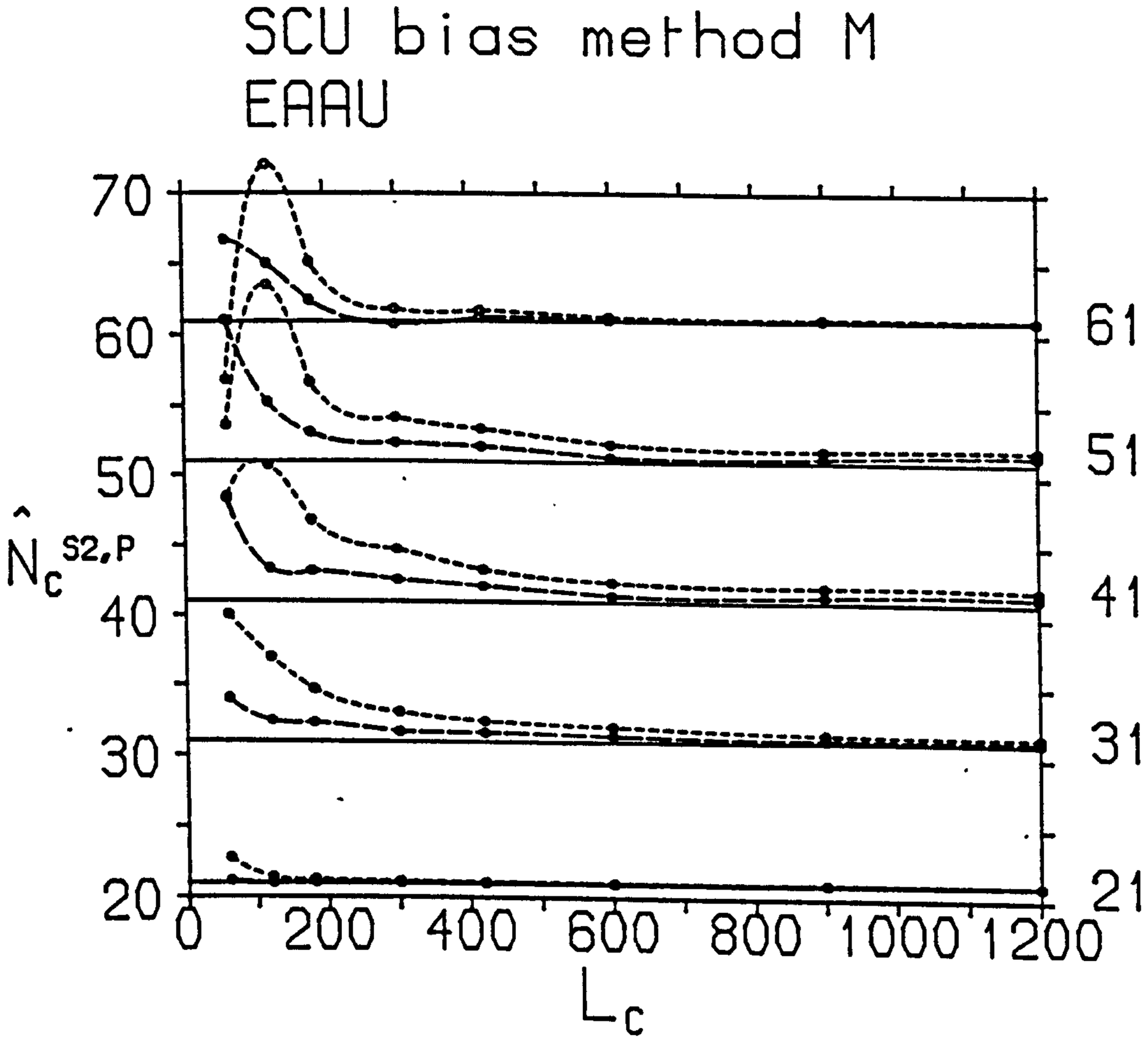

SCU bias method M
ECU
70
60
50
40
30
20
61
51
41
31
21
$\hat{N}_c^{S2,P}$
0
200
400
600
800
1000
1200
L_c

L

Opposite page 149

Figure 4.15

Comparison of estimators. $\hat{N}_c^p$ is identified as those lines with a longer dash. See pages 133–134 and 147 for further details.


SCU bias method M
EAAU
70
60
50
40
30
20
61
51
41
31
21
0
200
400
600
800
1000
1200
$\hat{N}_c^{S2,P}$
L_c

L

The range of $\hat{N}_c^p$ extends from 26.2 to 126.4, although the next highest figure is 73.3. There is good agreement between expression level and degree of SCU bias. The two categories of highly expressed genes (mainly ribosomal proteins) "effectively use" between 26 and 40 codons (i.e. the value of $\hat{N}_c^p$). Moderately expressed genes and regulatory genes "use" between 40 and 50 codons. Transposon and plasmid genes are relatively unbiased and "use" between 50 and 73 codons. Only four genes "use" more than 61 codons. The existence of $\hat{N}_c^p$ values greater than the number of sense codons is due to SCU bias being less than expected according to a uniform model.

There is one anomolous gene: rpoA, a highly expressed gene, has an extremely high value of 126.4. Closer examination of this gene reveals that four of the five four-codon families are used so rarely that the condition in equation (4.25) applies, thus inflating the contribution from this SF-type. However the three six-codon families exhibit considerable bias and are not used rarely. The size of this gene (L_c = 60) makes estimation of SCU bias difficult, but the problem is made worse by the biased amino-acid composition.

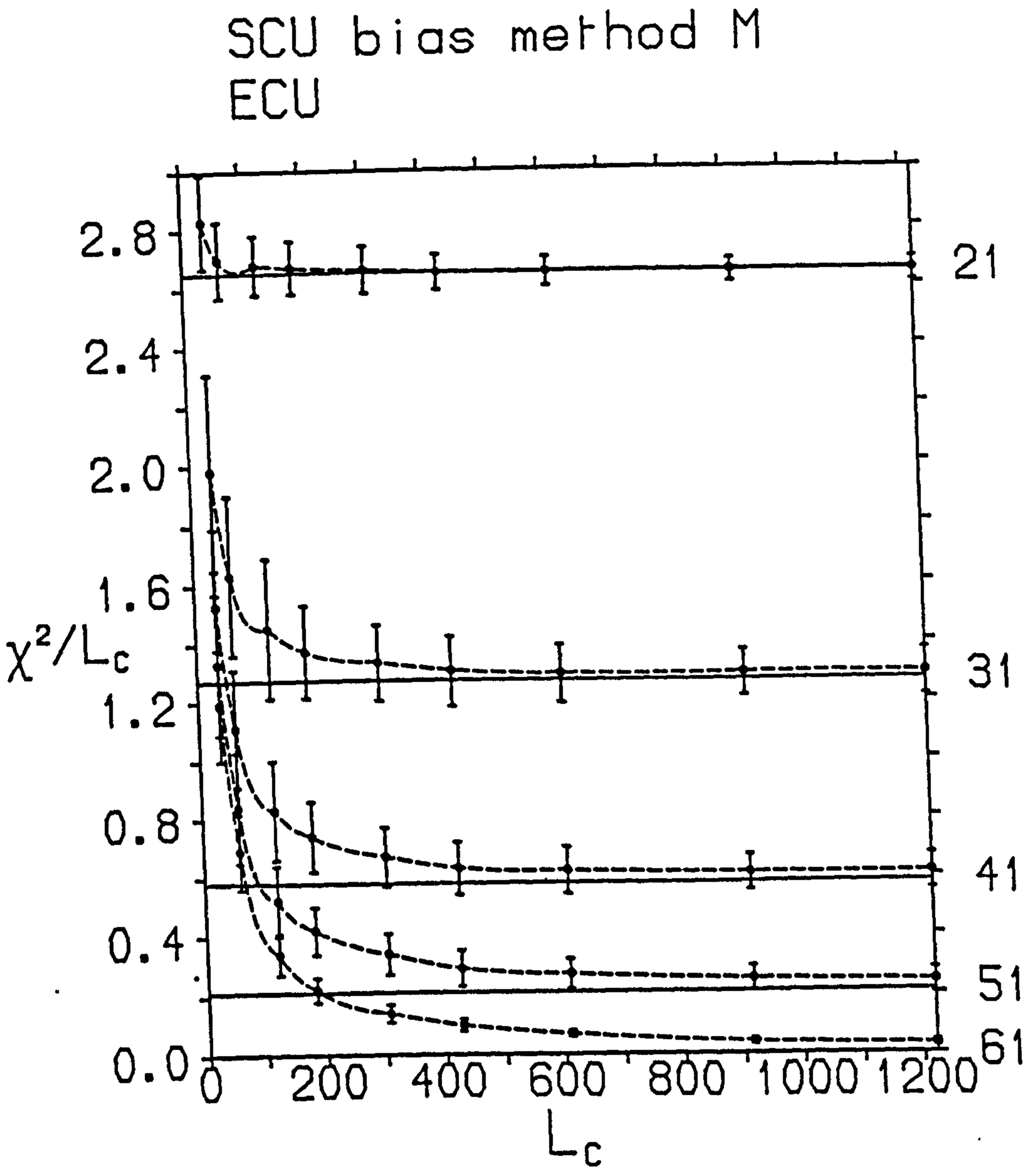
The χ^2 statistic/L_c, CPB, and $\hat{N}_c^p$ values are based on SCU reference pattern H_0: I.e. they are distance measures from an unbiased SCU pattern. The other four measures (CBI, CPS, f_{op} and CAI) are based on SCU reference patterns of the H_1 type. The CBI and f_{op} measures use the subset of 22 codons that tend to be used in *E.coli* highly-expressed genes as a reference pattern. The CPS and CAI measures use a reference SCU pattern prepared by pooling codon usage data from known highly-expressed genes.

If a large proportion of intra-specific variation in SCU bias is due to the correlation between level of expression and SCU bias, then the two types of bias measure will yield a similar ranking of genes. Table 4.5 confirms this to be true for *E.coli* genes. There is some evidence that the χ^2/L_c measure overestimates SCU bias: this observation was confirmed by rerunning the simulation program (see figures 4.16 and 4.17). The CPB measure, although a distance measure in multinomial standard deviations, is only quoted for relatively long genes thus precluding any significant effect of gene length.

Opposite page 151

Figure 4.16

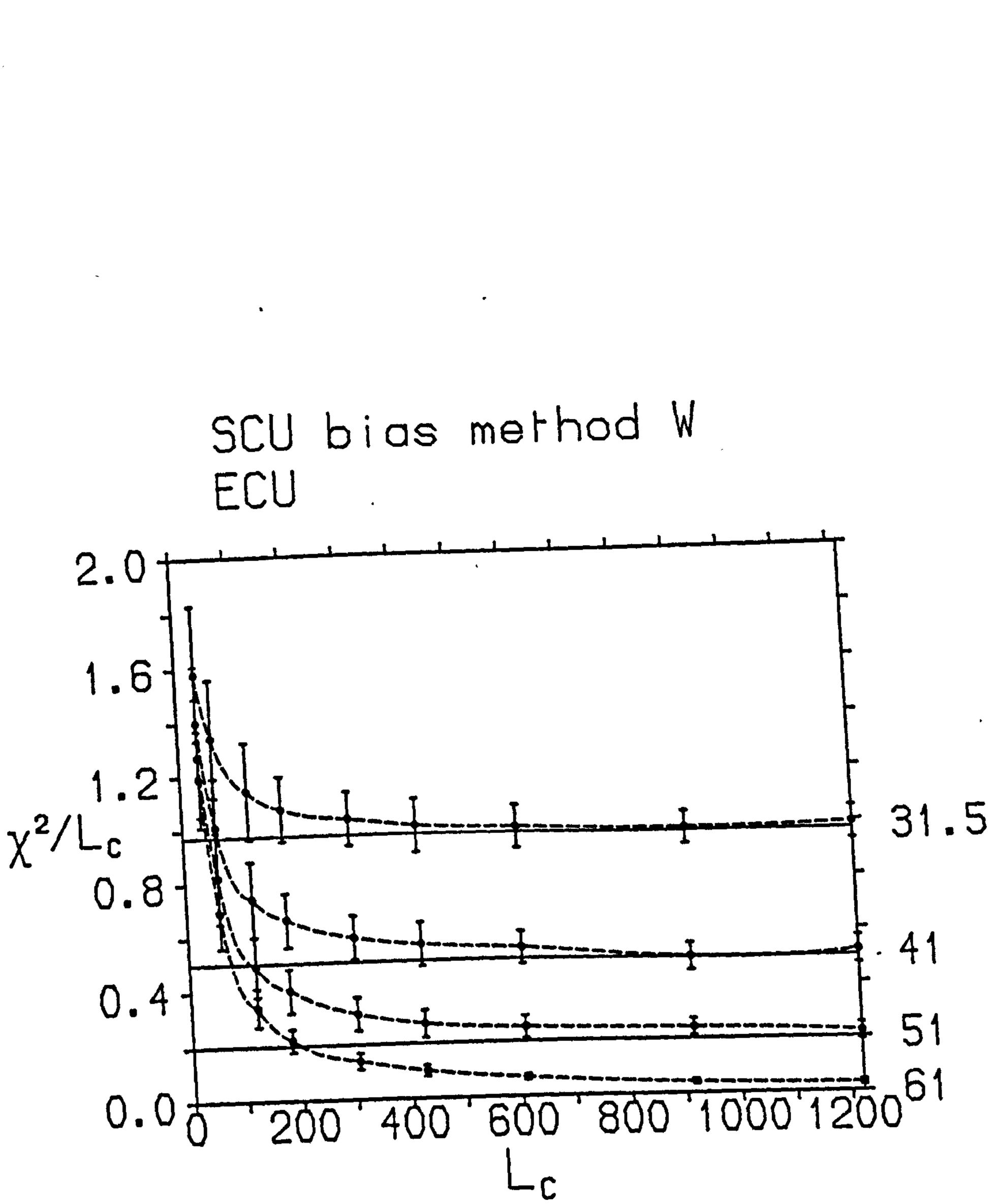
The behaviour of the χ^2/L_c measure of codon usage bias as a function of gene length L_c. See page 150 for discussion.


SCU bias method M
ECU
2.8
2.4
2.0
1.6
1.2
0.8
0.4
0.0
χ^2/L_c
0
200
400
600
800
1000
1200
L_c
21
31
41
51
61

Opposite page 152

Figure 4.17

The behaviour of the χ^2/L_c measure of codon usage bias as a function of gene length L_c. See page 150 for discussion.


SCU bias method W
ECU
2.0
1.6
1.2
0.8
0.4
0.0
χ^2/L_c
0
200
400
600
800
1000
1200
L_c
31.5
41
51
61

L

Yeast codon usage.

Fifty Yeast genes are listed in table 4.6. The basic information (class, gene, L_c, χ^2/L_c, and $\hat{N}_c^p$) is again supplemented with CBI and f_{op} values. However, instead of CPS and CPB values, the P2 statistic and "tRNA", the correlation between tRNA abundance and anticodon usage, is provided (data from Sharp *et al.* 1986). The classification of genes is shown in the first two columns of the table. "Plas" indicates a plasmid gene. The A-B-C classification refers to three groupings produced by a cluster analysis of the synonymous codon usage of 110 yeast genes by Sharp *et al.* (1986): A = high SCU bias, B = low bias and C represents a separate small group of outliers (mainly short genes). The high bias A group contained many known highly-expressed genes, implying that this may be a feature of all the genes in the group.

The range of $\hat{N}_c^p$ observed is from 22.6 to 63.1 codons. The most extremely biased yeast genes are more biased than those of *E.coli*: this point has already been noted by Bennetzen & Hall (1982). The tRNA values, while on average higher in group A, do not agree in detail with the ranking produced by the other bias measures. This is due to this correlation of tRNA abundance data and anticodon usage including information on amino-acid composition and excluding variation in SCU bias due to differential usage of codons that are recognised by the same tRNA. In contrast, the simple P2 statistic, which is based only on proportional usage of certain pairs of codons recognised by the same tRNA species, was in broad agreement with the CBI, χ^2/L_c, and $\hat{N}_c^p$ values.

One of the five C group outlier genes, the ribosomal protein L46 gene, is the most biased of the fifty genes studied. Although this estimate must be treated with caution due to the short gene length, a ribosomal protein gene would be expected to exhibit strong SCU bias. The cluster analysis of Sharp *et al.* (1986), while producing two main clusters consistent with two groups differing in expression level and SCU bias, did classify another ribosomal protein gene, S33, in the B group. This gene, again short, has a low $\hat{N}_c^p$ value (37.8) as expected for this class of gene.

The splitting of yeast genes into two distinct groups is not an obvious conclusion from their $\hat{N}_c^p$ values. There is a suggestion of a cut-off at about N_c^p = 35 codons. The sample of fifty genes may be too small, or the an SCU bias measure based on a H_1 reference pattern may be required to detect this effect, for example the CAI (Sharp & Li 1987).

Opposite page 155

Table 4.1

MSE figures for simulation M/ECU (see text for discussion). Each column represents a constant gene length L_c. The first three are MSE figures for the three estimators at a given level of SCU bias (i.e. N_c = 61). The rest of the table is ordered in the same way.

Table 4.2

Relative efficiency figures for the two main estimators studied (see text for discussion). Each column represents a constant gene length L_c. Each row represents a constant level of SCU bias.

Table 4.1 : MSE figures for Simulation M/ECU.

L_c	61	122	183	305	427	610	915	1220	
$\hat{N}_c^{s1}$	446.8	188.0	101.1	46.8	26.2	13.7	6.6	3.8	
$\hat{N}_c^{s2}$	128.6	45.4	12.1	2.8	1.4	0.8	0.4	0.3	N_c=61
$\hat{N}_c^{p}$	363.7	23.8	8.3	2.8	1.5	0.8	0.4	0.3	
$\hat{N}_c^{s1}$	180.1	65.7	31.4	14.6	7.4	4.9	2.7	1.9	
$\hat{N}_c^{s2}$	142.6	68.9	25.3	9.3	7.1	4.0	2.2	2.0	N_c=51
$\hat{N}_c^{p}$	295.1	30.2	13.5	6.7	5.6	3.6	2.1	1.8	
$\hat{N}_c^{s1}$	43.9	14.1	6.2	3.8	2.7	2.0	1.5	1.2	
$\hat{N}_c^{s2}$	102.1	77.9	27.8	11.2	7.9	4.6	2.3	1.7	N_c=41
$\hat{N}_c^{p}$	196.9	25.8	10.7	6.1	4.8	3.2	1.8	1.4	
$\hat{N}_c^{s1}$	7.7	4.5	2.6	1.8	1.5	1.1	0.8	0.6	
$\hat{N}_c^{s2}$	35.5	40.8	15.6	5.7	3.6	2.1	1.1	0.9	N_c=31
$\hat{N}_c^{p}$	48.1	10.8	4.8	2.7	2.0	1.4	0.8	0.7	
$\hat{N}_c^{s1}$	0.7	0.4	0.3	0.2	0.1	0.1	0.1	0.1	
$\hat{N}_c^{s2}$	1.0	1.2	0.6	0.2	0.2	0.1	0.1	0.1	N_c=21
$\hat{N}_c^{p}$	0.9	0.4	0.3	0.1	0.1	0.1	0.1	0.1	

Table 4.2 : Relative Efficiency of $\hat{N}_c^{p}$ and $\hat{N}_c^{s2}$ for Simulation M/ECU.

L_c	61	122	183	305	427	610	915	1220	
Rel.Eff.	2.41	1.24	1.26	1.19	1.18	1.12	1.03	1.12	N_c=61
	2.29	1.14	1.24	1.14	1.11	1.14	1.14	1.06	N_c=51
	2.46	0.97	1.12	1.09	1.04	1.08	1.02	1.02	N_c=41
	1.28	0.74	0.81	0.95	1.00	1.02	1.00	0.95	N_c=31
	0.96	0.40	0.59	0.74	0.76	0.88	0.92	0.91	N_c=21

Opposite page 156

Tables 4.3 and 4.4

These are arranged similarly to Table 4.1. See text for discussion.

Table 4.3 : Bias – simulation M/ECU.

L_c	61	122	183	305	427	610	915	1220	
$\hat{N}_c^{s1}$	-21.0	-13.6	-10.0	-6.8	-5.0	-3.6	-2.5	-1.9	
$\hat{N}_c^{s2}$	3.3	5.3	2.4	0.7	0.4	0.2	0.1	0.1	N_c=61
$\hat{N}_c^{p}$	9.0	1.3	0.7	0.1	0.1	0.1	0.1	0.1	
$\hat{N}_c^{s1}$	-13.1	-7.8	-5.2	-3.4	-2.1	-1.6	-1.0	-0.7	
$\hat{N}_c^{s2}$	5.9	6.8	4.0	1.9	1.7	1.0	0.6	0.6	N_c=51
$\hat{N}_c^{p}$	6.9	2.0	1.2	0.5	0.9	0.4	0.2	0.3	
$\hat{N}_c^{s1}$	-6.2	-2.8	-1.6	-0.8	-0.3	-0.1	-0.2	-0.1	
$\hat{N}_c^{s2}$	6.0	7.6	4.5	2.5	2.1	1.5	0.8	0.6	N_c=41
$\hat{N}_c^{p}$	5.9	2.4	1.5	0.9	1.0	0.7	0.3	0.3	
$\hat{N}_c^{s1}$	-1.4	-0.1	0.3	0.3	0.3	0.3	0.1	0.1	
$\hat{N}_c^{s2}$	2.2	5.4	3.3	1.8	1.3	1.0	0.5	0.4	N_c=31
$\hat{N}_c^{p}$	2.9	1.4	1.0	0.6	0.5	0.4	0.1	0.1	
$\hat{N}_c^{s1}$	0.0	0.1	0.0	0.1	0.1	0.0	0.0	0.0	
$\hat{N}_c^{s2}$	-0.2	0.6	0.2	0.2	0.1	0.1	0.0	0.1	N_c=21
$\hat{N}_c^{p}$	0.1	0.1	-0.0	0.0	0.0	0.0	0.0	0.0	

Table 4.4 : MSE figures for simulation M/EAAU.

L_c	60	120	180	300	420	600	900	1200	
$\hat{N}_c^{s1}$	591.0	269.7	162.2	83.3	46.0	25.9	12.5	7.8	
$\hat{N}_c^{s2}$	50.4	200.8	45.3	6.9	4.7	1.7	0.9	0.5	N_c=61
$\hat{N}_c^{p}$	280.7	77.8	31.3	7.3	5.0	1.8	0.9	0.5	
$\hat{N}_c^{s1}$	248.4	88.7	48.2	19.6	10.5	6.8	4.1	2.3	
$\hat{N}_c^{s2}$	49.9	244.8	57.5	19.5	11.4	5.0	3.7	2.9	N_c=51
$\hat{N}_c^{p}$	445.9	84.5	28.0	12.5	8.3	4.2	3.1	2.5	
$\hat{N}_c^{s1}$	63.9	19.3	10.2	4.2	3.5	2.3	2.0	1.4	
$\hat{N}_c^{s2}$	86.1	137.8	55.4	21.6	10.4	4.7	3.4	2.5	N_c=41
$\hat{N}_c^{p}$	158.5	30.2	22.5	10.4	6.5	2.9	2.4	1.8	
$\hat{N}_c^{s1}$	7.7	4.3	4.3	2.3	2.0	1.0	0.8	0.7	
$\hat{N}_c^{s2}$	105.4	55.3	22.8	7.7	4.5	2.2	1.2	0.9	N_c=31
$\hat{N}_c^{p}$	41.8	11.7	8.8	3.3	2.6	1.1	0.9	0.7	
$\hat{N}_c^{s1}$	0.8	0.3	0.3	0.2	0.1	0.1	0.1	0.0	
$\hat{N}_c^{s2}$	7.5	0.7	0.5	0.2	0.1	0.1	0.1	0.0	N_c=21
$\hat{N}_c^{p}$	0.9	0.3	0.3	0.2	0.1	0.1	0.1	0.0	

Opposite page 157

Table 4.5

See pages 147, 150 and 154 for discussion.

Table 4.5: SCU bias in 50 E.coli genes.

Class	Gene	L_c	G+C	χ^2/L_c	$\hat{N}_c^p$	CBI	CPS	CPB	f_{op}	CAI
VH	rpl L	120	46.39	1.37	26.21	0.84	1.57		0.96	
VH	rps A	555	51.17	1.15	29.17		1.66			
VH	lpp	77	51.52	1.53	29.37	0.84	1.76		0.98	0.849
VH	rps B	240	51.67	1.07	30.16					
VH	rpl A	233	51.93	1.08	30.75		1.64			
H	unc A	507	53.32	0.96	31.90		1.49		0.87	0.665
VH	tsf	282	50.12	0.97	32.11				0.91	
H	rpo C	1406	53.86	1.08	32.84		1.59			
VH	rec A	352	54.36	0.89	33.35		1.38	18.6	0.85	
VH	rpl K	141	52.72	0.99	33.46		1.50			
VH	omp A	345	53.91	1.07	33.95	0.78	1.60	19.4	0.92	
H	rpo B	1341	52.67	0.82	35.77	0.53	1.46	17.8	0.83	
VH	rpl J	164	52.85	0.96	36.34		1.41			
VH	rps L	123	53.93	1.22	36.53		1.55			
VH	rps G	80	49.17	1.32	36.98					
H	unc E	78	54.70	1.33	37.37				0.86	0.583
VH	rps T	86	47.29	1.05	37.76					
VH	rps J	102	56.21	1.16	39.99					
MOD	unc G	286	53.73	0.61	40.01					
MOD	asn A	329	56.74	0.62	41.04		1.02			
MOD	mal K	51	48.37	0.83	41.72					
MOD	lam B	395	50.97	0.47	41.78		1.19		0.78	
MOD	pur F	503	53.08	0.51	41.91		0.99		0.69	
MOD	ndh	433	53.12	0.56	42.35				0.70	
MOD	trp D	530	56.35	0.52	42.45				0.65	
VH	fus A	92	52.90	0.81	43.74					
H	ala S	875	53.71	0.50	43.82					
TP	colE1 imm	112	28.87	0.65	43.93					
MOD	trp B	396	54.80	0.50	44.25		0.99	12.8	0.70	
MOD	unc B	270	50.49	0.72	45.27				0.68	0.400
MOD	trp E	519	54.40	0.45	45.95				0.66	
REG	ara C	291	53.15	0.43	46.12		0.67	11.5	0.54	
MOD	lac Y	416	46.47	0.41	47.19		0.87	12.4	0.61	
MOD	unc F	155	53.12	0.56	47.28				0.73	
MOD	ilv G	299	55.30	0.47	47.41		0.91			
REG	lac I	359	56.45	0.45	47.50	0.18	0.76	10.6	0.63	0.296
MOD	thr A	819	53.07	0.30	48.14		0.79	10.9	0.61	
REG	lex A	201	52.24	0.55	48.38		0.98		0.67	
MOD	trp C	449	53.08	0.30	48.65		0.79		0.62	
MOD	trp A	267	54.06	0.44	49.40		0.84	10.7	0.63	
MOD	fol	158	53.16	0.49	50.53					
TP	Tn3 tr.	1014	51.71	0.30	50.92					
MOD	unc H	176	52.27	0.41	51.68				0.62	0.374
TP	Tn9 amp	218	45.11	0.27	55.60					
REG	trp R	107	54.52	0.61	57.43		0.62		0.56	0.267
TP	Tn3 B lact	285	49.59	0.17	59.64					
TP	Plasmid fol	77	56.28	0.48	63.01					
TP	PACYC KM-R	109	46.48	0.38	65.77					
TP	Tn1681	71	30.99	0.63	73.32					
H	rpo A	60	53.33	1.04	126.43					

Table 4.6: SCU bias in 50 yeast genes.

Class	Gene	L_c	G+C	χ^2/L_c	$\hat{N}_c^p$	CBI	P2	tRNA	f_{op}
C	r-pro L46	52	39.74	1.87	22.62	0.93	1.00	0.13	
A	GA-3-PDH 1	331	46.53	1.63	24.10	0.99	0.86	0.81	0.99
A	enolase B	438	47.41	1.66	24.94	0.96	0.86	0.75	0.98
A	enolase A	438	46.35	1.64	25.50	0.93	0.85	0.78	0.98
A	Tef 1	459	44.44	1.50	25.94	0.93	0.83	0.73	
A	GA-3-PDH 3	331	46.63	1.46	26.35	0.94	0.81	0.78	0.99
A	Adh 1	349	48.42	1.38	26.47	0.91	0.79	0.74	0.92
A	r-pro 51B	137	39.42	1.56	27.94	0.87	0.86	0.68	
A	r-pro L34	114	43.27	1.32	28.58	0.84	0.79	0.57	
A	r-pro 59	138	46.14	1.67	29.35	0.88	0.79	0.63	
A	r-pro L16	175	42.29	1.34	29.46	0.83	0.80	0.79	
A	actin	376	43.79	1.13	29.57	0.82	0.80	0.83	0.91
A	Hxk 2	486	41.84	1.05	30.98	0.73	0.72	0.84	
A	histone 3	137	43.80	1.44	31.93	0.77	0.68	0.61	
A	histone 2B1	132	40.91	1.30	32.29	0.77	0.77	0.62	0.88
A	hsp 90	710	39.06	0.87	33.13	0.66	0.68	0.64	
B	ATP 2	313	41.96	0.81	37.11	0.50	0.53	0.78	
B	r-pro S33	68	45.10	1.10	37.78	0.63	0.79	0.31	
B	Pho5	468	42.24	0.59	38.07	0.47	0.66	0.67	
A	Ubiquitin	382	41.45	0.70	38.11	0.50	0.68	0.71	
A	histone 2B2	132	41.92	1.06	38.68	0.71	0.63	0.65	
B	Mating F. a	166	44.58	0.67	38.91	0.34	0.62	0.39	
B	Trp 5	708	43.22	0.56	40.32	0.45	0.65	0.76	0.57
B	PPR 2	129	39.28	0.60	41.64	0.30	0.59	0.53	
A	iso-1-cyt.C	110	42.42	0.66	41.83	0.47	0.63	0.51	0.72
B	Gal 7	185	38.56	0.59	41.90	0.20	0.48	0.75	
B	Cpa 2	1119	38.96	0.39	42.95	0.30	0.53	0.78	
A	Porin	284	45.31	0.61	44.03	0.50	0.65	0.74	
C	Mat A1	149	30.65	0.48	44.94	0.00	0.32	0.38	
B	Mn SOD	234	49.43	0.40	45.03	0.34	0.63	0.60	
B	His 4	800	42.21	0.39	45.45	0.37	0.59	0.63	
B	Car 1	334	47.50	0.39	47.27	0.34	0.63	0.78	
B	Rad 3	779	37.10	0.31	48.28	0.10	0.41	0.59	
B	Gal 10	446	40.51	0.29	49.06	0.14	0.43	0.73	
B	Ura 3	268	43.03	0.42	49.33	0.21	0.51	0.71	
C	UCCR	148	46.17	0.32	49.60	0.23	0.55	0.29	
B	Trp 2	529	39.57	0.28	50.00	0.17	0.48	0.69	
B	Gal 1	529	41.46	0.24	50.83	0.20	0.46	0.67	
B	Rad 52	505	43.04	0.23	50.90	0.09	0.46	0.58	
B Plas	2μ able	424	37.03	0.23	51.15	-0.06	0.47	0.50	
B	Rad 10	196	37.07	0.38	51.32	0.06	0.47	0.53	
C	YP2	207	45.73	0.37	51.33	0.25	0.67	0.47	
B Plas	PKT1	317	45.53	0.21	52.17	-0.03	0.48	0.53	
B Plas	2μ charlie	297	41.75	0.26	52.17	0.06	0.49	0.55	
B	CBP 1	655	37.10	0.17	54.01	0.11	0.46	0.57	
B Plas	2μ baker	374	39.84	0.18	55.82	-0.05	0.38	0.48	
B	Gal 4	882	40.82	0.11	56.78	0.04	0.44	0.63	
B	Mat A2	120	37.22	0.37	61.80	-0.04	0.45	0.25	
B	CDC8	217	38.25	0.23	63.05	0.12	0.43	0.56	
C	Cup 1 Cu chel.	62	43.55	0.66	66.82	0.11	0.62	0.12	

4.8. Discussion.

The effective number of codons $\hat{N}_c^p$ is a simple, easily computable measure of SCU bias that provides a conceptually obvious description of a coding sequence. The absence or very rare usage of several amino-acids does decrease the accuracy of $\hat{N}_c^p$, but these conditions are rarely encountered in real genes. This simple distance measure from uniform synonymous codon usage allows the quick estimation of SCU bias from the codon usage of the particular gene without the requirement for other information (e.g. compilations of codon usage for highly expressed genes).

$\hat{N}_c^p$ is therefore of great use for organisms where little is known of codon usage patterns and few genes have been sequenced.

$\hat{N}_c^p$ also allows comparisons between genes from different species. Measures that are based on comparisons with known species-specific bias patterns cannot be used simply for inter-specific comparisons. This is because such measures are relative to a known bias pattern, whereas methods like the one used here yield absolute measures of SCU bias.

Within bacterial and fungal species, there is little variation in the G+C content of genes (Sueoka 1961a). This is confirmed in the analysis of the codon usage of fifty *E.coli* and fifty yeast genes (see Tables 4.5. and 4.6). The use of $\hat{N}_c^p$ for intra-specific comparisons of SCU bias is not significantly affected by G+C variation. Inter-specific comparisons between genes from these two species will be influenced by the 10 per cent difference in overall species G+C content. This effect might be expected to show up in the $\hat{N}_c^p$ values for weakly biased yeast genes being less biased than those of *E.coli*. There is no evidence of this. The highly- expressed genes of *E.coli* and yeast exhibit strong SCU bias patterns that are not explicable in terms of their differences in G+C content: indeed, the G+C difference is less in this class of genes.

However, much of the variation in the codon usage patterns of vertebrate genes is due to variation in G+C content (Bernardi *et al.* 1985). $\hat{N}_c^p$ will thus yield low values (high SCU bias) for genes with extreme G+C content in the third codon position. If only that component of SCU bias independent of G+C

content is required, a method based on knowledge of base composition could be used.

CHAPTER 5
GENERAL DISCUSSION AND CONCLUSIONS.

5.1. Patterns of Codon Usage.

The study of codon usage bias has advanced enormously since 1981 when the work contained in this thesis began. The correspondence analysis detailed in chapter one was carried out before the full importance of DNA base composition as a factor was realised. G+C content has turned out to be the most important nucleic acid property w.r.t. codon usage and the early work by Sueoka (1961a,b) on G+C patterns has assumed a new relevance (Sueoka 1986). Other features of nucleic acid molecules (e.g. DNA/RNA, single/double stranded molecule) appear less important. The strand asymmetry of mammalian mitochondrial genomes (due to G+T imbalance) is another base compositional trend observed affecting codon usage patterns. There is evidence of more complex patterns than can be simply explained by G+C content: The doublet CpG is not underrepresented in high G+C vertebrate DNA (Adams *et al.* 1987); Bernardi & Bernardi (1986) note that G+C differences between eukaryotic viral genes are mainly due to differences in C alone.

Mutation has been suggested by many authors (e.g. Sueoka (1962,1986), Sharp & Li (1987), Jukes & Bhushan (1987), Filipski (1987)) as the cause of the patterns of DNA base composition observed. Unicellular organisms have a relatively small variance in coding G+C content compared to vertebrate organisms. Sueoka's simple mutation pressure model (Sueoka 1962) predicts that each unicellular organism has a mean G+C content and that genes from such an organism will reflect this mean G+C content and show a variance in G+C content appropriate for a binomial proportion. However, it now appears that the SCU patterns of some unicellular genes are under selection (Sharp & Li 1987).

The SCU patterns of highly-expressed unicellular genes appear to be subject to selection, but also exhibit a small range in G+C content. The SCU patterns of three bacterial species, *E.coli*, *S.typhimurium* (Ikemura 1985a), and *B.subtilis* (Shields & Sharp 1987), and of yeast (Ikemura 1985a) show only a small variation in G+C content. The correspondence analysis plots showed

that most of the *E.coli* and yeast codon usage patterns were independent of G+C content. This suggests that the unidirectional trends in SCU bias seen in unicellular organisms (Sharp & Li 1987) may be constrained. A consistent choice of G+C rich/poor optimal codons would result in G+C rich/poor highly expressed genes. This may be the explanation of the apparent "counterbalance effect" observed by Wada & Suyama (1986).

The existence of DNA regions of differing mean G+C content in warm-blooded vertebrate genomes may be due to local differences in mutation pressure (Filipski 1987) as opposed to the action of selection on each nucleotide (Bernardi & Bernardi 1986). The observed high G+C between-gene variation observed in human sequence data may therefore reflect both the mosaic nature of vertebrate genomic structure and G+C variation in the local DNA region. Adams *et al.* (1987) have noted that 150-200 bp fluctuations in G+C content may be related to nucleosome positioning.

Biased mutation pressure appears to influence amino-acid composition, in addition to SCU, in human genes (see chapter 3), warm-blooded vertebrate genes (Bernardi & Bernardi 1986), and in mammalian mitochondria (Jukes & Bhushan 1987). This suggests that the relationship of codon usage and other factors should be the focus of more study. For example, a comparison of amino-acid usage in highly and lowly expressed unicellular genes might reveal more about the relationship between codon usage patterns and tRNA abundance. The lack of independence of amino-acid usage and SCU bias in vertebrates also suggests that the relationship between silent and replacement changes should be further studied. Lipman & Wilbur (1985) noted a small reduction in SCU bias in regions of vertebrate genes that were not highly conserved.

About 40 per cent of the variation in codon usage revealed by the correspondence analysis was due to G+C content, the base compositional bias of mammalian mitochondrial genes, highly expressed genes of *E.coli* and yeast, and differences between the universal and mitochondrial codes. No other obvious general patterns were discovered. This result, along with the increased understanding of unicellular (Sharp & Li 1987) and vertebrate (Bernardi & Bernardi 1986; Filipski 1987) codon usage patterns, suggests that a pooled analysis of codon usage data from a wide range of genomes is not

appropriate.

The "Genome Hypothesis" (Grantham 1980a) is thus confirmed. Indeed, the absence of general patterns (other than G+C content) means that the low dimensional plots produced by correspondence analysis of pooled data from many genomes are of limited use. Genes that are not extreme in G+C content, and that do not come from unicellular or mitochondrial genomes are poorly displayed on the first four axes of the correspondence analysis. This suggests that cluster analysis would be more appropriate for the analysis of codon usage data from a range of genomes, given the considerable substructure in the data. Further discussion of statistical methodology is contained in section 5.3.1.

5.2. Measures of Synonymous Codon Usage Bias.

Synonymous codon usage bias patterns in vertebrate nuclear genomes and certain mitochondrial genomes (mammal, drosophila, and yeast) appears to be largely explained by DNA base composition. However intra-specific SCU patterns in unicellular organisms appear to independent of the base composition of the particular gene. The measure of SCU bias developed in chapter 4 will be useful in identifying highly-expressed genes in unicellular genomes which do not have extreme G+C content.

5.3. Theoretical Analysis of Codon Usage Patterns.

5.3.1. Data Exploration.

The correspondence analysis detailed in this thesis was used to carry out an exploratory data analysis. The presence of substructure in the data analysed, mainly due to species-specific codon usage patterns, may mean that the choice of data exploration technique may have to be reconsidered. Emphasis could be placed on species differences in codon choice or on within-species variation.

The distance measure used in correspondence analysis is proportional to a χ^2 metric. The χ^2 distance is appropriate for contingency tables, but

consideration of such a measure w.r.t. evolutionary distances may be a fruitful line of research. The use of other metrics may reveal more about evolutionary forces. The measure of SCU bias developed in chapter 4 utilised the analogy between allele frequencies at separate loci and the frequency of synonymous codons in different amino-acids. After this work was completed, it was realised that the same analogy had been used by Nei & Tajima (1981) to develop a measure of DNA diversity. This "nucleon diversity" is essentially the same as the expression developed in chapter 4 for the effective number of codons used by a single amino acid. The above analogy could be used to adapt other measures of genetic distance developed for allele frequency data for use in the study of codon usage pattern variation. Care must be exercised in interpreting such results given the knowledge that not all genes are subject to the same selection and/or mutation pressure.

The next stage would be to use a model appropriate to data in the form of a contingency table to carry out hypothesis testing. An obvious model is the Generalized Linear Model (McCullagh & Nelder 1983). A statistical programming language (GLIM) (Baker & Nelder 1978) has been produced to facilitate the use of models of this type.

5.3.2. Population Genetics Models.

Recent papers (e.g. Li 1987, Bulmer 1987) have produced theoretical models of codon usage based on population genetics theory. Li (1987) has developed Kimura's (1981, 1983) simple model of stabilising selection acting on synonymous codon usage patterns. Bulmer (1987) investigated the coevolution of codon usage and tRNA abundance, under the assumption that selection acts on translation speed. If tRNA abundance and codon usage are poorly adapted to each other then the average time to translate codons will be high due to the ribosome waiting for tRNA molecules to be recharged with the respective amino-acids. Selection pressure is therefore inversely proportional to tRNA abundance.

Bulmer's model predicts that highly expressed genes will have a considerable effect on the coevolution of tRNA abundance and codon usage; that, in *E.coli* and yeast, weak selection pressure and high population size will overcome mutation pressure and thus produce biased codon usage pattern.

Bulmer also notes that multicellular organisms are likely to be subject to lower selection pressures at the level of translation and have smaller population sizes, and their codon usage patterns may therefore be dominated by mutation pressure.

A component of evolutionary models of codon usage relates to a sub-model of the translational system. The coevolution model of Bulmer (1987) assumes that selection pressure is inversely proportional to tRNA abundance. Holm (1986) has elaborated on the model of Gouy & Grantham (1980), and applied it to *E.coli*. She concludes that translational accuracy is the subject of optimisation, not translational speed. A detailed review of codon usage and translation has been published by de Boer & Castelein (1986). However, a detailed translational model is not required for population genetics models of codon usage.

The theoretical contributions of Li (1987), and Bulmer (1987) provide a solid evolutionary framework in which to view the phenomenon of SCU bias. It is likely that models devised for hypothesis testing will use this theoretical framework.

5.4. Future Research.

Most of the current research on codon usage models focusses on unicellular and vertebrate genomes. These evolutionary groups differ in many ways, but they serve as useful model systems. It is likely that chloroplast codon usage patterns will have much in common with unicellular patterns: codon usage and tRNA abundance appear to be coevolving in chloroplasts (Pfitzinger *et al.* 1987). There are however large gaps in our "natural history" of codon usage. Comprehensive studies of plant nuclear and mitochondrial genomes, and of invertebrate genomes are only now becoming possible due to the continuing increase in sequence data.

I. Appendix: Genes Analysed by Correspondence Analysis.

Eukaryotic Nuclear subgroup: EN_1 (N=35; all *H.sapiens*).

No.	source	gene/product	genome
1	EMBL02:HSAGL1	α-globin	*H.sapiens* .
2	EMBL02:HSALB1	serum albumin	.
3	EMBL02:HSATRP	α-1-antitrypsin	.
4	EMBL02:HSBGL3	β-globin	.
5	EMBL02:HSDGL1	δ-globin	.
6	EMBL02:HSEGL1	ε-globin	.
7	EMBL02:HSGGL2	γ_a-globin	.
8	EMBL02:HSGGL4	γ_g-globin	.
9	EMBL02:HSGONA	chorionic gonadotrophin	.
10	EMBL02:HSGROW1	growth hormone	.
11	EMBL02:HSHLA1	class1 transplant'n antigen	.
12	EMBL02:HSIFD1	α-interferon LeIF-J	.
13	EMBL02:HSIFD2	α-interferon LeIF-C1	.
14	EMBL02:HSIFR10	α-interferon LeIF-D	.
15	EMBL02:HSIFR12	α-interferon LeIF-F	.
16	EMBL02:HSIFR13	α-interferon LeIF-G	.
17	EMBL02:HSIFR14	α-interferon LeIF-H	.
18	EMBL02:HSIFR15	γ-interferon	.
19	EMBL02:HSIFR7	α-interferon LeIF-A	.
20	EMBL02:HSIFR8	α-interferon LeIF-B	.
21	EMBL02:HSIFR9	α-interferon LeIF-C	.
22	EMBL02:HSIG03	immunoglobulin γ-2	.
23	EMBL02:HSIGK1	C-terminal κ-immunoglobulin	.
24	EMBL02:HSIGK2	invariant region: κ-Ig	.
25	EMBL02:HSIGK3	variable region: κ-Ig	.
26	EMBL02:HSIGM1-3	immunoglobulin μ	.
27	EMBL02:HSINSU	preproinsulin	.
28	EMBL02:HSLACT	prolactin	.
29	EMBL02:HSMGLO	β-2 microglobulin	.
30	EMBL02:HSOPIO	proopiomelanocortin	.
31	EMBL02:HSSOMA	somatomammotropin	.
32	EMBL02:HSTHIO	metallothionein	.
33	EMBL02:HSTHYR	preproparathyroid hormone	.
34	EMBL02:HSENK1,-2	preproenkephalin	.
35	EMBL02:HSIFD4	β-1 interferon	.

Eukaryotic Nuclear subgroup: EN_2 (N=9; all *G.gallus*).

No.	source	gene/product	genome
36	EMBL02:GGAGL1	α-globin	*G.gallus*
37	EMBL02:GGATUB	α-tubulin	.
38	EMBL02:GGBTUB	β-tubulin	.
39	EMBL02:GGCO10	pro-α-2(I) collagen	.
40	EMBL02:GGCOL9	pro-α-1(I) collagen	.
41	EMBL02:GGH2AX	histone H2a	.
42	EMBL02:GGVLDI	low density lipoprotein II	.
43	EMBL02:GGALB2	ovalbumin	.
44	EMBL02:GGINS1,-2	insulin	.

Eukaryotic Nuclear subgroup: EN_4 (N=7; all *Xenopus laevis*).

No.	source	gene/product	genome
45	EMBL02:XLBGL2	β-globin	*X.laevis*
46	EMBL02:XLRIB1	ribosomal protein L1	.
47	EMBL02:XLRIB2	ribosomal protein L14	.
48	EMBL02:XLRIB3	ribosomal protein L32	.
49	EMBL02:XLRIB6	ribosomal protein S19	.
50	EMBL02:XLRIB4	ribosomal protein S1	.
51	EMBL02:XLRIB5	ribosomal protein S8	.

Eukaryotic Nuclear subgroup: EN_3 (N=10; six Pisces species).

No.	source	gene/product	genome
52	EMBL02:LASOM1	somatostatin I	*L.americanus*
53	EMBL02:LASOM2	somatostatin II	*L.americanus*
54	EMBL05:LAINSU	preproinsulin	*L.americanus*
55	EMBL05:LAGLUC	preproglucagen	*L.americanus*
56	EMBL05:TCACH1	acetylcholine receptor γ s/u	*L.americanus*
57	EMBL02:MGINSU	preinsulin	*M.glutinosa*
58	EMBL05:SSINSU	preproinsulin	Siberian salmon
59	EMBL05:SGHIS01	histone H2a	*Salmo gairdneri*
60	EMBL05:SGHIS01	histone H3	*Salmo gairdneri*
61	EMBL05:IPSOM1	somatostatin-14	*I.punctatus*

Eukaryotic Nuclear subgroup: EN_6 (N=5; all *Psammechinus miliaris*).

No.	source	gene/product	genome
62	EMBL02:PMHIS7	histone H4	*P.miliaris*
63	EMBL02:PMHIS7	histone H2b	.
64	EMBL02:PMHIS7	histone H3	.
65	EMBL02:PMHIS7	histone H2a	.
66	EMBL02:PMHIS7	histone H1	.

Eukaryotic Nuclear subgroup: EN_5 (N=11; all *Drosophila melanogaster*).

No.	source	gene/product	genome
67	EMBL05:DMCUT2	cuticle protein gene II	*D.melanogaster*
68	EMBL05:DMCUT2	cuticle protein gene III	.
69	EMBL05:DMCUT2	cuticle protein gene IV	.
70	EMBL02:DMADHS	ADH-s	.
71	EMBL05:DMCUT3	cuticle protein D	.
72	EMBL05:DMHS08	hsp22	.
73	EMBL05:DMHS09	hsp23	.
74	EMBL05:DMHS10	hsp26	.
75	EMBL05:DMHS11	hsp27	.
76	EMBL02:DMYOLK	vitellogenin	.
77	EMBL02:DMHSP2	hsp70	.

Eukaryotic Nuclear subgroup: EN7 (N=8; six Protozoan species).

No.	source	gene/product	genome
78	EMBL05:ACAC01	actin I	*A.castelani*
79	EMBL05:TBGP01	surface BC glycoprotein	*T.brucei*
80	EMBL05:TTHI01	histone H4-I gene	*T.thermophila*
81	Helftenbein(1985)	α-tubulin	*S.lemnae*
82	Preer *et al.* (1985)	A-antigen Dehyd)	*P.tetraurelia*
83	Horowitz & Gorovsky (1985)	histone H3I	*T.thermophila*
84	Horowitz & Gorovsky (1985)	histone H3II	*T.thermophila*
85	Caron & Meyer (1985)	G-antigen	*P.primaurelia*

Eukaryotic Nuclear subgroup: EN8 (N=16; all *Saccharomyces Cerevisiae*).

No.	source	gene/product	genome
86	EMBL02:SCACT1	actin	*S.cerevisiae*
87	EMBL02:SCADHI	ADH(isozyme I)	
88	EMBL02:SCCYT1	iso-1-cytochrome C	
89	EMBL02:SCCYT2	iso-2-cytochrome C	
90	EMBL02:SCGAP1	pgap49 gene (glyc.3.phos. Dehyd)	
91	EMBL02:SCHIS1	histone H2B1	
92	EMBL02:SCHIS2	histone H2B2	
93	EMBL02:SCTRP1	trp1	
94	EMBL02:SCTRP5A	trp5	
95	EMBL02:SCMAT1A	gene cassette:A mating type	
96	EMBL02:SCMAT2A	gene cassette:α1 mating type	
97	EMBL02:SCMAT2A	gene cassette:α2 mating type	
98	EMBL02:SCGAP2	pgap63 gene (glyc.3. phos.dehyd)	
99	EMBL02:SCH2A1	histone H2A1	
100	EMBL02:SCH2A2	histone H2A2	
101	EMBL02:SCHIS4A	his4 (histidine metabolism)	

Eukaryotic Nuclear subgroup: ENg (N=7; five Spermatophyte species).

No.	source	gene/product	genome
102	EMBL05:PHCHAL	chalcone synthetase	Parsley
103	EMBL02:PSCH01	small s/u chloroplast carboxylase	*P.sativum*
104	EMBL05:GMRUBP	ribulose-1,5-bisphos--phate carboxylase	*Glycine max*
105	EMBL05:GMGLO2	leghemoglobin gene (Lbc)	*Glycine max*
106	EMBL02:PVPHAS	phaseolin	*P.vulgaris*
107	EMBL05:ZMZE01	zein	*Zea mays*
108	EMBL02:GMACTI	actin	*Glycine max*

Prokaryotic subgroup: P_1 (N=50; all *E.coli* chromosomal genes).

No.	source	gene/product	genome
109	EMBL02:ECATPX	putative unc protein	*E.coli*
110	EMBL02:ECATPX	uncB	.
111	EMBL02:ECATPX	uncE	.
112	EMBL02:ECATPX	uncF	.
113	EMBL02:ECATPX	uncH	.
114	EMBL02:ECATPY	uncG	.
115	EMBL02:ECATPY	uncD	.
116	EMBL02:ECATPY	uncC	.
117	EMBL02:ECLACY	lacY	.
118	EMBL02:ECLAMBA	lamB	.
119	EMBL02:ECMALK	malK	.
120	EMBL02:ECMALX	malT	.
121	EMBL02:ECNDHX	ndh	.
122	EMBL02:ECOMPA	ompA	.
123	EMBL02:ECPHOA	phoA	.
124	EMBL02:ECRPOBC	rplK	.
125	EMBL02:ECRPOBC	rplA	.
126	EMBL02:ECRPOBC	rplJ	.
127	EMBL02:ECROPBC	rplL	.
128	EMBL02:ECRPOBC	rpoB	.
129	EMBL02:ECROPBC	rpoC	.
130	EMBL02:ECTRPX	trpE	.
131	EMBL02:ECTRPX	trpD	.
132	EMBL02:ECTRPX	trpC	.
133	EMBL02:ECTRPX	trpB	.
134	EMBL02:ECTRPX	trpA	.
135	EMBL02:ECRPSB	rpsB	.
136	EMBL02:ECRPSB	tsf	.
137	EMBL02:ECRPSL	rpsJ	.
138	EMBL02:ECRPSL	rplC	.
139	EMBL02:ECRSPA	rpsA	.
140	EMBL02:ECS4AS	rpoA	.
141	EMBL02:ECS4AS	rpsD	.
142	EMBL02:ECSTR1	rpsL	.
143	EMBL02:ECSTR2	fusA	.
144	EMBL02:ECSTR1,-2	rpsG	.
145	EMBL02:ECILVX	ilvG	.
146	EMBL02:ECALAS	alaS	.
147	EMBL02:ECARAC	araC	.
148	EMBL02:ECASNA	asnA	.
149	EMBL02:ECATPXA	uncA	.
150	EMBL02:ECFOLX	fol	.
151	EMBL02:ECLACI	laci	.
152	EMBL02:ECLEXX	lexA	.
153	EMBL02:ECLPPX	lpp	.
154	EMBL02:ECPURF	purF	.
155	EMBL02:ECRECA	recA	.
156	EMBL02:ECRPST	rpsT	.
157	EMBL02:ECTHRA	thrA	.
158	EMBL02:ECTRPR	trpR	.

Prokaryotic subgroup: P_2 (N=8; all *E.coli* plasmid/ transposon genes).

No.	source	gene/product	genome
159	EMBL02:EC5388	*E.coli* plasmid-assoc'd dihydrofolate reductase	*E.coli* (plasmid).
160	EMBL02:ECCOL1	*E.coli* plasmid colE1 imm (immunity) gene	. .
161	EMBL02:ECKAMR	*E.coli* plasmid PACYC KM-R. (kanamycin resistance)	.
162	EMBL02:ISTN16	*E.coli* Transposon Tn1681 heat-stable (ST) toxin	*E.coli* (transposons).
163	EMBL02:ISTN3X	*E.coli* Transposon Tn3 repressor gene	.
164	EMBL02:ISTN3X	*E.coli* Transposon Tn3 beta-lactamase	.
165	EMBL02:ISTN3X	*E.coli* Transposon Tn3 transposase	.
166	EMBL02:ISTN9X	*E.coli* Transposon Tn9 amp (chloramphenicol resistence)	. . .

Prokaryotic subgroup: P_3 (N=11; all *S.typhimurium* genes).

No.	source	gene/product	genome
167	EMBL02:STARGT	argT	*S.typhimurium*
168	EMBL02:STHINX	HIN	.
169	EMBL05:STHISX	hisJ	.
170	EMBL05:STHISX	hisQ	.
171	EMBL05:STHISX	putative his operon gene	.
172	EMBL05:STHISX	hisP	.
173	EMBL02:STTRPA	trpA	.
174	EMBL02:STRTPB	trpB	.
175	EMBL05:STTRPE	trpE	.
176	Erickson *et al.* (1985)	dnaG	.
177	Erickson *et al.* (1985)	rpoD	.

Prokaryotic subgroup: P_4 (N=8; four Rhizopoda species).

No.	source	gene/product	genome
178	EMBL02:RLMODABC	nodA protein	*R.legumin-osarum*
179	EMBL02:RLMODABC	nodB protein	.
180	EMBL02:RLMODABC	nodC protein	.
181	EMBL05:RJNIFDK	nifdk Operon:dinitrogenaseα s/u	*R.japonicum* .
182	EMBL02:RMNIFH	nifH (nitrogenase reductase)	*R.meliloti* .
183	EMBL05:ATNOPA	*A.tumef.* plasmid nopaline synthase	*A.tumefaciens* .
184	EMBL05:ATOCTO	A.tumef plasmid octopine synthase	.
185	EMBL05:ATTMS2	*A.tumef.* plasmid pTiA6 tms transcript 2 locus	. .

Prokaryotic subgroup: P_5 (N=4; all *Anabaena* 7120).

No.	source	gene/product	genome
186	EMBL05:A7NIFX	nifH (nitrogenase reductase)	*Anabaena* 7120
187	EMBL05:A7NIFX	nifD (dinitrogenase α s/u)	.
188	EMBL05:ANGLNA	glna (glutamine synthetase)	.
189	EMBL05:ANRUBP	ss-gene (small s/u ribulose-1,5-biphosphate carboxylase/oxygenase)	.

Prokaryotic subgroup: P_6 (N=19; all *Bacillus subtilis*).

No.	source	gene/product	genome
190	EMBL05:BSAMYL	amylase	*B.subtilis*
191	Yoshikawa *et al.*(1986)	0.5kb gene (formerly spoOH)	.
192	Ogasawara (1985)	dnaD	.
193	Ogasawara (1985)	dnaN	.
194	Ogasawara (1985)	recF	.
195	Ogasawara (1985)	gyrB	.
196	Ogasawara (1985)	gyrA	.
197	Ogasawara (1985)	rpmH	.
198	Ogasawara (1985)	dnaE	.
199	Ogasawara (1985)	trpE	.
200	Ogasawara (1985)	trpD	.
201	Ogasawara (1985)	aprE	.
202	Ogasawara (1985)	0.3kb	.
203	Ogasawara (1985)	spoOB	.
204	Ogasawara (1985)	spoIIG	.
205	Ogasawara (1985)	trpC	.
206	Ogasawara (1985)	spoOA	.
207	Ogasawara (1985)	nprE	.
208	Ogasawara (1985)	rpmA	.

Prokaryotic subgroup: P_7 (N=4; all *Mycoplasma capricolum*).

No.	source	gene/product	genome
209	Yamao *et al.* (1985)	ribosomal protein S3	*M.capricolum*
210	Yamao *et al.* (1985)	ribosomal protein L16	.
211	Muto *et al.* (1984)	ribosomal protein S8	.
212	Muto *et al.* (1984)	ribosomal protein L18	.

Eukaryotic Organelle subgroup: EO_2 (N=5; all *S.Cerevisiae* mtDNA).

No.	source	gene/product	genome
213	EMBL05:MISC18	ATPase proteolipid	Yeast mtDNA
214	EMBL02:MISC02	ATPase	.
215	EMBL02:MISCX1	COII	.
216	EMBL02:MISC05	cytb	.
217	EMBL02:MISC13	COI	.

Eukaryotic Organelle subgroup: EO_1 (N=8; all *H.sapiens* mtDNA).

No.	source	gene/product	genome
218	EMBL02:MIHSXX	URF1	*H.sapiens* mtDNA
219	EMBL02:MIHSXX	COI	.
220	EMBL02:MIHSXX	COII	.
221	EMBL02:MIHSXX	URFA6L	.
222	EMBL02:MIHSXX	ATPase6	.
223	EMBL02:MIHSXX	COIII	.
224	EMBL02:MIHSXX	URF3	.
225	EMBL02:MIHSXX	cytb	.

Eukaryotic Organelle subgroup: EO_7 (N=8; all *B.taurus* mtDNA).

No.	source	gene/product	genome
226	EMBL02:MIBTXX	URF1	*B.taurus* mtDNA
227	EMBL02:MIBTXX	COI	.
228	EMBL02:MIBTXX	COII	.
229	EMBL02:MIBTXX	URFA6L	.
230	EMBL02:MIBTXX	ATPase6	.
231	EMBL02:MIBTXX	COIII	.
232	EMBL02:MIBTXX	URF3	.
233	EMBL02:MIBTXX	cytb	.

Eukaryotic Organelle subgroup: EO_6 (N=8; all *M.musculus* mtDNA).

No.	source	gene/product	genome
234	EMBL02:MITOMM	URF1	*M.musculus* mtDNA
235	EMBL02:MITOMM	COI	.
236	EMBL02:MITOMM	COII	.
237	EMBL02:MITOMM	URFA6L	.
238	EMBL02:MITOMM	ATPase6	.
239	EMBL02:MITOMM	COIII	.
240	EMBL02:MITOMM	URF3	.
241	EMBL02:MITOMM	cytb	.

Eukaryotic Organelle subgroup: EO_5 (N=4; all *D.melanogaster* mtDNA).

No.	source	gene/product	genome
242	GenBank:DROMTM1	ATPase6	*D.mel.* mtDNA
243	GenBank:DROMTM1	COIII	.
244	GenBank:DROMTM1	COI	.
245	GenBank:DROMTM1	COII	.

Eukaryotic Organelle subgroup: EO_3 (N=2; 2sp. Spermatophyte mtDNA).

No.	source	gene/product	genome
246	EMBL02:MIZMCO	COII	*Z.mays*
247	EMBL05:MIOV01	COII	*O.villaricea*

Eukaryotic Organelle subgroup: EO_4 (N=9; cpDNA: eight sp.).

No.	source	gene/product	genome
248	EMBL05:CHGM01	psbA	*G.max*
249	EMBL05:CHHV01	atpB (ATPase B s/u)	*H.vulgare*
250	EMBL05:ECTUFA	rps12	*E.glacilis*
251	EMBL05:ECTUFA	rps7	.
252	EMBL05:CHNT02	ATPase α s/u	*N.tabacum*
253	EMBL02:CHSORC	rubpase	*S.oleracea*
254	EMBL02:CHZM02	rubpcase	*Z.mays*
255	EMBL02:SAPSII	pre-M(r)32000 PSII protein	*S.alba*
256	EMBL02:SOCPA1	P(860)chlorophyll-a apoprotein gene	*S.oleracea*

Bacteriophage subgroup: B_1 (N=46; all Bacteriophage λ).

No.	source	gene/product	genome
257	EMBL02:LAMBDA	Nu1	λ
258	EMBL02:LAMBDA	A	.
259	EMBL02:LAMBDA	W	.
260	EMBL02:LAMBDA	B	.
261	EMBL02:LAMBDA	C	.
262	EMBL02:LAMBDA	D	.
263	EMBL02:LAMBDA	E	.
264	EMBL02:LAMBDA	F1	.
265	EMBL02:LAMBDA	F11	.
266	EMBL02:LAMBDA	Z	.
267	EMBL02:LAMBDA	U	.
268	EMBL02:LAMBDA	V	.
269	EMBL02:LAMBDA	G	.
270	EMBL02:LAMBDA	T	.
271	EMBL02:LAMBDA	H	.
272	EMBL02:LABDAM	M	.
273	EMBL02:LAMBDA	L	.
274	EMBL02:LAMBDA	K	.
275	EMBL02:LAMBDA	I	.
276	EMBL02:LAMBDA	J	.
277	EMBL02:LAMBDA	cro	.

Bacteriophage subgroup: B_1 contd. (N=46; all Bacteriophage λ).

278	EMBL02:LAMBDA	cII	λ
279	EMBL02:LAMBDA	O	.
280	EMBL02:LAMBDA	P	.
281	EMBL02:LAMBDA	ren	.
282	EMBL02:LAMBDA	Q	.
283	EMBL02:LAMBDA	R	.
284	EMBL02:LAMBDA	Rz	.
285	EMBL02:LAMBDA	Ea47	.
286	EMBL02:LAMBDA	Ea31	.
287	EMBL02:LAMBDA	Ea59	.
288	EMBL02:LAMBBDA	int	.
289	EMBL02:LAMBBDA	xis	.
290	EMBL02:LAMBDA	Ea8.5	.
291	EMBL02:LAMBDA	Ea22	.
292	EMBL02:LAMBBDA	exo	.
293	EMBL02:LAMBDA	B	.
294	EMBL02:LAMBDA	Y	.
295	EMBL02:LAMBDA	kil	.
296	EMBL02:LAMBDA	CIII	.
297	EMBL02:LAMBDA	ssb	.
298	EMBL02:LAMBDA	ral	.
299	EMBL02:LAMBDA	N	.
300	EMBL02:LAMBDA	rexB	.
301	EMBL02:LAMBDA	rexA	.
302	EMBL02:LAMBDA	cI	.

Eukaryotic Viral subgroup: EV_1 (N=3; all Epstein-Barr Virus).

No.	source	gene/product	genome
303	EMBL05:EBV	RRED	EBV
304	EMBL05:EBV	GLYC	.
305	EMBL05:EBV	DPO1	.

Bacteriophage subgroup: B_6 (N=5; all Bacteriophage ϕ29).

No.	source	gene/product	genome
306	EMBL02:POP29A	Gene 6	ϕ29
307	EMBL02:POP29B	Gene 2	.
308	EMBL02:POP29B	Gene 3	.
309	Garvey *et al.* (1985a)	Gene 16	.
310	Garvey *et al.* (1985a)	Gene 17	.

Bacteriophage subgroup: B_7 (N=4; all Bacteriophage P22).

No.	source	gene/product	genome
311	EMBL02:POP22E	erf	P22
312	EMBL02:POP22X	c2 repressor	.
313	Rennell & Poteete (1985)	gene 13 (lysis)	.
314	Rennell & Poteete (1985)	gene 19 (lysis)	.

Bacteriophage subgroup: B_2 (N=41; all Bacteriophage T7).

No.	source	gene/product	genome
315	EMBL02:PODOT7	0.3	T7
316	EMBL02:PODOT7	0.4	.
317	EMBL02:PODOT7	0.6a	.
318	EMBL02:PODOT7	0.7	.
319	Moffat *et al.* (1984)	1	.
320	EMBL02:PODOT7	1.2	.
321	EMBL02:PODOT7	1.3	.
322	EMBL02:PODOT7	1.4	.
323	EMBL02:PODOT7	1.6	.
324	EMBL02:PODOT7	1.7	.
325	EMBL02:PODOT7	2	.
326	EMBL02:PODOT7	2.5	.
327	EMBL02:PODOT7	2.8	.
328	EMBL02:PODOT7	3	.
329	EMBL02:PODOT7	3.5	.
330	EMBL02:PODOT7	3.8	.
331	EMBL02:PODOT7	4a	.
332	EMBL02:PODOT7	4.3	.
333	EMBL02:PODOT7	4.5	.
334	EMBL02:PODOT7	4.7	.
335	EMBL02:PODOT7	5	.
336	EMBL02:PODOT7	5.3	.
337	EMBL02:PODOT7	5.5	.
338	EMBL02:PODOT7	5.7	.
339	EMBL02:PODOT7	6	.
340	EMBL02:PODOT7	7	.
341	EMBL02:PODOT7	7.3	.
342	EMBL02:PODOT7	7.7	.
343	EMBL02:PODOT7	7.8	.
344	EMBL02:PODOT7	7.9	.
345	EMBL02:PODOT7	10a	.
346	EMBL02:PODOT7	11	.

Bacteriophage subgroup: B_2 contd. (N=41; all Bacteriophage T7).

347	EMBL02:PODOT7	12	T7
348	EMBL02:PODOT7	13	.
349	EMBL02:PODOT7	14	.
350	EMBL02:PODOT7	15	.
351	EMBL02:PODOT7	16	.
352	EMBL02:PODOT7	17	.
353	EMBL02:PODOT7	17.5	.
354	EMBL02:PODOT7	18	.
355	EMBL02:PODOT7	19	.

Bacteriophage subgroup: B_3 (N=11; all Bacteriophage T4).

No.	source	gene/product	genome
356	EMBL02:MYT4TL	gene36	T4
357	EMBL02:MYT4TL	gene37	.
358	EMBL05:MYT4DENV	endonuclease V	.
359	EMBL05:MYT4LIG	ligase	.
360	EMBL05:MYT4LY	lisozyme	.
361	EMBL05:MYT4RL	RNA ligase	.
362	EMBL05:MYT4TR	g45 protein	.
363	EMBL05:MYT4TR	g44 protein	.
364	EMBL05:MYT4TR	g62 protein	.
365	EMBL05:MYT4TR	regA	.
366	EMBL05:MYT4TR	g43 protein	.

Eukaryotic Viral subgroup: EV_8 (N=5; all Cauliflower Mosaic Virus).

No.	source	gene/product	genome
367	EMBL02:CAMVG1	I	Cauli.M.V.
368	EMBL02:CAMVG1	II	.
369	EMBL02:CAMVG1	III	.
370	EMBL02:CAMVG1	IV	.
371	EMBL02:CAMVG1	V	.

Eukaryotic Viral subgroup: EV_4 (N=2; all Mouse Minute Virus).

No.	source	gene/product	genome
372	EMBL05:PAMVM2	urf from small genome	M.V.M.
373	EMBL05:PAMVM2	urf from small genome	.

Bacteriophage subgroup: B_4 (N=5; two Levivirus species).

No.	source	gene/product	genome
374	EMBL02:LEMS2X	A protein	MS2
375	EMBL02:LEMS2X	coat protein	.
376	EMBL02:LEMS2X	L protein	.
377	EMBL02:LEMS2X	replicase beta s/u	.
378	EMBL02:LEQBET	coat protein	Qβ

Eukaryotic Viral subgroup: EV_9 (N=3; three Plant Mosaic virus sp.).

No.	source	gene/product	genome
379	EMBL02:TOTMV3	30k protein	Tobacco M.V.
380	EMBL02:TYTYM1	coat protein	Turnip Yellow M.V.
381	EMBL02:ALALM5	coat protein	Alfalfa M.V.

Eukaryotic Viral subgroup: EV_3 (N=5; all SV40).

No.	source	gene/product	genome
382	EMBL02:SV40XX	VP1	SV40
383	EMBL02:SV40XX	VP2	.
384	EMBL02:SV40XX	VP3	.
385	EMBL02:SV40XX	small t	.
386	EMBL02:SV40XX	large t (part1+part2)	.

Eukaryotic Viral subgroup: EV_2 (N=22; all Adenovirus type 2).

No.	source	gene/product	genome
387	EMBL05:AD2	e1b 20.5k	Adenovirus type 2
388	EMBL05:AD2	e1b 57k	.
389	EMBL05:AD2	ix protein	.
390	EMBL05:AD2	e3 19k	.
391	EMBL05:AD2	e3 11.6k	.
392	EMBL05:AD2	13.6k	.
393	EMBL05:AD2	52/55k	.
394	EMBL05:AD2	iiia protein	.
395	EMBL05:AD2	penton protein	.
396	EMBL05:AD2	pro-vii protein	.
397	EMBL05:AD2	pv protein	.
398	EMBL05:AD2	pvi protein	.
399	EMBL05:AD2	hexon protein	.
400	EMBL05:AD2	23k protein	.
401	EMBL05:AD2	100k protein	.
402	EMBL05:AD2	pviii protein	.
403	EMBL05:AD2	fiber protein	.
404	EMBL05:AD2	e4 11k	.
405	EMBL05:AD2	DNA polymerase	.
406	EMBL05:AD2	bellet protein	.
407	EMBL05:AD2	DNA binding protein	.
408	EMBL05:AD2	33k protein	.

Bacteriophage subgroup: B_5 (N=9; all Bacteriophage ϕX174).

No.	source	gene/product	genome
409	EMBL02:PHIX174	gene C (phage maturation)	ϕX174
410	EMBL02:PHIX174	gene D (phage assembly)	.
411	EMBL02:PHIX174	gene E (host cell lysis)	.
412	EMBL02:PHIX174	gene J (core protein)	.
413	EMBL02:PHIX174	gene F (major coat prot.)	.
414	EMBL02:PHIX174	gene G (major spike prot)	.
415	EMBL02:PHIX174	gene H (minor spike prot)	.
416	EMBL02:PHIX174	gene A (DNA replication)	.
417	EMBL02:PHIX174	gene B (capsid morphogenesis)	.

Eukaryotic Viral subgroup: EV_5 (N=6; all Polio virus).

No.	source	gene/product	genome
418	EMBL02:POLIOS1	VP4	Polio virus
419	EMBL02:POLIOS1	VP2	Sabin 1 str.
420	EMBL02:POLIOS1	VP3	.
421	EMBL02:POLIOS1	VP1	.
422	EMBL02:POLIOS1	3b	.
423	EMBL02:POLIOS1	1b	.

Eukaryotic Viral subgroup: EV_6 (N=3; three retroviruses).

No.	source	gene/product	genome
424	EMBL02:NOAMVX	transforming gene	avian myelo--blastosis virus
425	EMBL02:REASVY	protein p90gag-yes	avian sarcoma virus
426	EMBL02:REMSVX	gag polyprotein	murine sarcoma .

Eukaryotic Viral subgroup: EV_7 (N=2; all Simian 11 rotavirus).

No.	source	gene/product	genome
427	EMBL05:REROT1	VP7 protein	SA11
428	EMBL05:ROSRV6	major inner capsid prot.	.

II. References.

1. Adams R.L.P., Davis T., Rinaldi A., Eason R. (1987) CpG deficiency, dinucleotide distributions and nucleosome positioning. Eur.J.Biochem. 165:107-115.

2. Adams J., Rothman E.D. (1982) Estimation of phylogenetic relationships from DNA restriction patterns and selection of endonuclease cleavage sites. Proc. Natl. Acad. Sci. USA 79:3560-3564.

3. Alff-Steinberger C. (1969) The genetic code and error transmission. Proc. Natl. Acad. Sci. USA 64:584-591.

4. Alonso S., Minty A., Bourlet Y., Buckingham M. (1986) Comparison of three actin-coding Sequences in the Mouse; Evolutionary relationships between the actin Genes of warm-blooded vertebrates. J.Mol.Evol. 23:11-22.

5. Aota S-I., Ikemura T. (1986) Diversity in G+C content at the third position of codons in vertebrate genes and its cause. Nucleic Acids Res. 14:6345-6355.

6. Argos P., Rossman M.G., Grau U.M., Zuber H., Frank G., Tratschin J.D. (1979) Thermal stability and protein structure. Biochemistry 18:5698-5703.

7. Baer R., Bankier A.T., Biggin M.D., Deininger P.L., Farrell P.J., Gibson T.J., Hatfull G., Hudson G.S., Satchwell S.C., Seguin C., Tuffnell P.S., Barrell B.G. (1984) DNA Sequence and expression of the B95-8 Epstein-Barr virus genome. Nature 310:207-211.

8. Baker R.J., Nelder J.A. (1978) The GLIM System, Release 3, Generalized Linear Interactive Modelling, Numerical Algorithms Group, Oxford.

9. Bennetzen J.L., Hall B.D. (1982) Codon selection in yeast. J.Biol.Chem. 257:3026-3031.

10. Bernardi G., Bernardi G. (1985) Codon usage and genome composition. J.Mol.Evol. 22:363-365.

11. Bernardi G., Bernardi G. (1986) Compositional constraints and genome evolution. J.Mol.Evol. 24:1-11.

12. Bernardi G., Olofsson B., Filipski J., Zerial M., Salinas J., Cuny G., Meunier-Rotival M., Rodier F. (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953-958.

13. Bibb M.J., Findlay P.R., Johnson M.W. (1984) The relationship between base composition and codon usage in bacterial

genes and its use for the simple and reliable identification of protein-coding regions. Gene 30:157-166.

14. Bird A.P. (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. 8:1499-1504.

15. Bird A.P. (1986) CpG-rich islands and the function of DNA methylation. Nature 321:209-213.

16. Bishop Y.M.M., Fienberg S.E., Holland P.W. (1975) Discrete multivariate analysis: Theory and practice. MIT Press, Cambridge, Mass.

17. Birnstiel M., Spiers J., Purdom I.F. (1972) Kinetic complexes of RNA molecules. J.Mol.Biol. 63:21-39.

18. Blaisdell B.E. (1983a) A prevalent persistent global nonrandomness that distinguishes coding and noncoding eukaryotic nuclear DNA sequences. J.Mol.Evol. 19:122-133.

19. Blaisdell B.E. (1983b) Choice of base at silent codon 3 is not selectively neutral in eukaryotic structural genes: It maintains excess short runs of weak and strong hydrogen bonding bases. J.Mol.Evol. 19:226-236.

20. Blake R.D., Hinds P.W. (1984) Analysis of the codon bias in *E.coli* sequences. J.Biomol.Struct.Dyn. 2:593-606.

21. Bonitz S.G., Bertani R., Coruzzi G., Li M., Macino G., Nobrega F.G., Nobrega M.P., Thalenfeld B.E., Tzagoloff A. (1980) Codon recognition rules in yeast mitochondria. Proc. Natl. Acad. Sci. USA 77:3167-3170.

22. Bossi L. (1983) Context effects: Translation of UAG codon by suppressor tRNA is affected by the sequence following UAG in the message. J.Mol.Biol. 164:73-87.

23. Bossi L., Roth J.R. (1980) The influence of codon context on genetic code translation. Nature 286:123-127.

24. Brown W.M. (1980) Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. Proc. Natl. Acad. Sci. USA 77:3605- 3609.

25. Brown W.M. (1981) Mechanisms of evolution in animal mitochondrial DNA. Ann. N. Y. Acad. Sci. 361:119-134.

26. Brown W.M. (1983) Evolution of animal mitochondrial DNA. Chapter 4 In: Nei M., Koehn RK (Eds.) Evolution of genes and proteins. Sinauer. Sunderland, Massachusetts.

27. Brown W.M. (1985) The mitochondrial genome of animals. In: MacIntyre R.J. (ed) Molecular evolutionary genetics. Plenum, New York, pp95-130.

28. Bulmer M. (1987) Coevolution of codon usage and tRNA abundance. Nature 325:728-730.

29. Caron F., Meyer E. (1985) Does *Paramecium primaurelia* use a different genetic code in its macronucleus? Nature 314:185-188.

30. Chevallier A., GareL J.P. (1979) Differential synthesis rates of tRNA species in the silk gland of *Bombyx mori* are required to promote tRNA adaptation to silk messages. Eur.J.Biochem. 124:477-482.

31. Clarke B.C. (1970) Darwinian evolution of proteins. Science 168:1009-1011.

32. Clarke C.H. (1982) The influence of DNA base ratios on intragenic mutation spectra in prokaryotes. J.Theor.Biol. 94:671-687.

33. Clarke C.H. (1983) Differential amino acid composition of proteins in AT- & GC- rich eubacteria. Speculations in Science and Technology 6:113-123.

34. Cox E.C. (1976) Bacterial mutator genes and the control of spontaneous mutation. Ann. Rev. Genet. 10:135-156.

35. Crick F.H.C., Brenner S., Klug A., Pieczenik G. (1976) Origins of Life 7:389-397.

36. Crow J.F., Kimura M. (1970) An introduction to population genetics theory. Harper & Row, New York.

37. Darland G., Brock T.D., Samsonoff W., Conti S.F. (1970) A thermophilic, acidophilic *Mycoplasma* isolated from a coal refuse pile. Science 170:1416-1418.

38. de Boer H.A., Castelein R.A. (1986) Biased codon usage: An exploration of its role in optimization of translation. In: From gene to protein; Steps dictating the maximal level of gene expression. Managing editor: Davis J., Editors: Reznikoff W.S., Gold L. Butterworths, Stoneham, Mass., p225-283.

39. Devereux J., Haeberli P., Smithies O. (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. 12:387-395.

40. Edelman G.M., Gallant J. (1977) Mistranslation in *E.coli.* Cell 10:131-137.

41. Eigen M., Schuster P. (1979) The hypercycle. Springer-Verlag, Berlin.

42. Elton R.A. (1973) The relationship of DNA base composition

and individual protein composition in micro-Organisms. J.Mol.Evol. 2:263-276.

43. EMBL Nucleotide Sequence Data Library (release 2: April 83), European Molecular Biology Laboratory, Postfach 10 22 09, D-6900 Heidelberg, FRG.

44. Erickson B.D., Burton Z.F., Watanabe K.K., Burgess R.R. (1985) Nucleotide sequence of the rpsU-dnaG-rpoD operon from *Salmonella typhimurium* and a comparison of this sequence with the homologous operon of *Escherichia coli* Gene 40:67-78.

45. Filipski J. (1987) Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. FEBS letters 217:184-186.

46. Fincham J.R.S. (1983) Genetics. Wright PSG, Bristol.

47. Foster P.L., Eisenstadt E., Cairns J. (1982) Random components in mutagenesis. Nature 299:365-367.

48. Freese E. (1962) On the evolution of base composition of DNA. J.Theor.Biol. 3:82-101.

49. Frommel C., Holzhutter H-G. (1985) An estimate on the effect of point mutation and natural selection on the rate of amino-acid replacement in proteins. J.Mol.Evol. 21:233-257.

50. Fox T.D. (1985) Diverged genetic codes in protozoans and a bacterium. Nature 314:132-133.

51. Garel J.P. (1976) Quantitative adaptation of isoacceptor tRNAs to mRNA codons of alanine, glycine and serine. Nature 260:805-806.

52. Garvey K.J., Saedi M.S., Ito J. (1985a) The complete sequence of *Bacillus* phage ϕ29 gene 16: a protein required for the genome encapsidation reaction. Gene 40:311-316.

53. Garvey K.J., Yoshikawa H., Ito J. (1985b) The complete sequence of the *Bacillus* phage ϕ29 right early region. Gene 40:301-309.

54. Gauss D.H., Sprinzl M. (1984a) Compilation of tRNA genes. Nucleic Acids Res. 12:r1-r57.

55. Gauss D.H., Sprinzl M. (1984b) Compilation of sequences of tRNA genes. Nucleic Acids Res. 12:r59-r131.

56. Gillham N., Boynton J.E., Harris E.H. (1985) Evolution of plastid DNA. In: T. Cavalier-Smith (ed) The Evolution of

genome size. Wiley, Chichester, UK.

57. Gittins R. (1985) Canonical Analysis: a review of applications in ecology. Springer-Verlag, Heidelberg.

58. Golding G.B., Strobeck C. (1982) Expected frequencies of codon use as a function of mutation rates and codon fitnesses. J.Mol.Evol. 18:379-386.

59. Goldman M.A., Holmquist G.P., Gray M.C., Caston L.A., Nag A. (1984) Replication timing of Genes and middle repetitive sequences. Science 224:686-692.

60. Golub G.H., Van Loan C.F. (1983) Matrix Computations. John Hopkins University Press, Baltimore, Maryland.

61. Gouy M., Gautier C. (1982) Codon usage in bacteria: Correlation with gene expressivity. Nucleic Acids Res. 10:7055-7074.

62. Grantham R., Gautier C., Gouy M. (1980a) Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res. 8:1893-1912.

63. Grantham R., Gautier C., Gouy M., Mercier R., and Pave A. (1980b) Codon catalog usage and the genome hypothesis. Nucl.Acids Res. 8: r49-r62.

64. Grantham R., Gautier C., Gouy M., Jacobzone M., Mercier R. (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9:r43-r74.

65. Grantham R., Greenland T., Louail S., Mouchiroud D., Prato J.L., Gouy M., Gautier C. (1985) Molecular evolution of viruses as seen by nucleic acid sequence study. Bull. Inst. Pasteur. 83:95-148.

66. Greenacre M.J. (1984) Theory and applications of correspondence analysis. Academic Press, London.

67. Greenacre M.J., Vrba E.S. (1984) Graphical display and interpretation of antelope census data in african wildlife areas, using correspondence analysis. Ecology 65:984-997.

68. Gribskov M., Devereux J., Burgess R.R. (1984) The codon preference plot: Graphic analysis of protein coding sequences and prediction of gene expression. Nucleic Acids Res. 12:539-549.

69. Grosjean H., Fiers W. (1982) Preferential codon usage in prokaryotic genes: The optimal codon-anticodon interaction energy and the selective codon usage in efficiently

expressed genes. Gene 18:199-209.

70. Guthrie C., Abelson J. (1982) Organisation and expression of tRNA genes in *Saccharomyces cerevisiae* In: The molecular biology of the yeast *Saccharomyces* metabolism and gene expression. (Strathern J.N., Jones E.W., Broach J.R., eds) Cold Spring Harbor Laboratory, New York.

71. Hastings K.E.M., Emerson C.P. (1983) Codon usage in muscle genes and liver genes. J.Mol.Evol. 19:214-218.

72. Hatlen L.E., Attardi G. (1971) Proportion of the HeLa cell genome complementary to tRNA and 5S RNA. J.Mol.Biol. 56:535-554.

73. Helftenbein E. (1985) Nucleotide sequence of a macronuclear DNA molecule coding for α-tubulin from the ciliate *Stylonychia lemnae* Special codon usage: TAA is not a translation termination codon. Nucleic Acids Res. 13:415-433.

74. Hill L.R. (1966) An index to deoxyribonucleic acid base compositions of bacterial species. J.Gen.Microbiol. 44:419-437.

75. Holm L. (1986) Codon usage and gene expression. Nucleic Acids Res. 14:3075-3087.

76. Horowitz S., Gorovsky M.A. (1985) An unusual genetic code in nuclear genes of *Tetrahymena* Proc. Natl. Acad. Res. USA 82:2452-2455.

77. Ikemura T. (1980) The frequency of codon usage in *E.coli* genes : Correlation with abundance of cognate tRNA. In: Genetics and evolution of RNA polymerase, tRNA and ribosomes. (ed. S. Osawa), University of Tokyo Press, Tokyo and Elsevier/North-Holland, pp519-523. Amsterdam.

78. Ikemura T. (1981a) Correlation between the abundance of *E.coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J.Mol.Biol. 146:1-21.

79. Ikemura T. (1981b) Correlation between the abundance of *E.coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E.coli* translational system. J.Mol.Biol. 151:389-409.

80. Ikemura T. (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in its protein genes. J.Mol.Biol. 158:573-597.

81. Ikemura T. (1985a) Codon usage and tRNA content in unicellular and multicellular organisms. Mol.Biol.Evol.

2:13-34.

82. Ikemura T. (1985b) Codon usage, tRNA content, and rate of synonymous substitution. In: Population Genetics and Molecular Evolution. (eds) Ohta T., Aoki K. pp385-406. Japan Sci. Soc. Press, Tokyo/ Springer-Verlag, Berlin.

83. Ikemura T., Ozeki H. (1983) Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents. Cold Spring Harbor Symp. Quant. Biol. 47:1087-1097.

84. Joliffe I.T. (1986) Principal component analysis. Springer-Verlag, Heidelberg.

85. Jukes T.H. (1983) Mitochondrial codes and evolution. Nature 301:19-20.

86. Jukes T.H., Bhushan V. (1986) Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. J.Mol.Evol. 24:39-44.

87. Jukes T.H., Osawa S., Muto A. (1987) Divergence and directional mutation pressures. Nature 325:668.

88. Kimura M. (1968) Evolutionary rate at the molecular level. Nature 217:624-6.

89. Kimura M. (1981) Possibility of extensive neutral evolution under stabilizing selection with special reference to non-random usage of synonymous codons. Proc. Natl. Acad. Sci. USA 78:5773-5777.

90. Kimura M. (1983) The neutral theory of molecular evolution. Cambridge University Press, London.

91. Kimura M., Crow J.F. (1964) The number of alleles that can be maintained in a finite population. Genetics 49:725-738.

92. King J.L., Jukes T.H. (1969) Non-Darwinian Evolution. Science 164:788-799.

93. Kreitman M. (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster.* Nature 304:412-417.

94. Kruger D.H., Bickle T.A. (1983) Bacteriophage survival: Multiple mechanisms for avoiding the deoxyribonucleic acid restriction systems of their hosts. Microbiol.Rev. 47:345-360.

95. Kunkel T.A., Schaaper R.M., Beckman R.A., Loeb L.A. (1981) On the fidelity of DNA replication: effect of the next nucleotide on proofreading. J.Biol.Chem. 256:9883-9889.

96. Lathe R. (1985) Synthetic probes deduced from amino-acid sequence data: theoretical and practical considerations. J:Mol.Biol. 183:1-12.

97. Lawrence C.W. (1982) Mutagenesis in *S.cerevisiae*. Advances in Genetics 21:173-254.

98. Lennon G.G., Nussinov R. (1984) Homonyms, synonyms and mutations of the sequence/structure vocabulary. J.Mol.Biol. 175:425-430.

99. Lewin B. (1983) Genes. Wiley, New York.

100. Li W-H. (1987) Models of nearly neutral mutation with particular implications for nonrandom usage of synonymous codons. J.Mol.Evol. 24:337-345.

101. Li W-H., Gojobori T., Nei M. (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292:237-239.

102. Li W-H., Luo C.C., Wu C.I. (1985) Evolution of DNA sequences. In: MacIntyre R.J. (ed) Molecular evolutionary genetics. Plenum, New York, pp1-94.

103. Lipman D.J., Maizel J. (1982) Comparative analysis of nucleic acids by their characteristic constraints. Nucleic Acids Res. 10:2723-2739.

104. Lipman D.J., Wilbur W.J. (1985) Interaction of silent and replacement changes in eukaryotic coding sequences. J.Mol.Evol. 21:161-167.

105. Maruyama T., Gojobori T., Aota S-I., Ikemura T. (1986) Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res. 14: r151-r197.

106. McCullagh P., Nelder J.A. (1983) Generalized linear models. Chapman and Hall, London.

107. McLachlan A.D., Staden R., Boswell D.R. (1984) A method for measuring the non-random bias of a codon usage table. Nucleic Acids Res. 12:9567-9575.

108. Modiano G., Battistuzzi G., Motulsky A.G. (1981) Nonrandom pattern of codon usage and of nucleotide substitution in human α- and β-globin genes: An evolutionary strategy reducing the rate of mutations with drastic effects? Proc. Natl. Acad. Sci. USA 78:1110-1114.

109. Moffat B.A., Dunn J.J., Studier F.W. (1984) Nucleotide sequence of the gene for bacteriophage T7 RNA polymerase. J.Mol.Biol. 173:265-269.

110. Muto A., Kawauchi Y., Yamao F., Osawa S. (1984)

Preferential use of A- and U- rich codons for *Mycoplasma capricolum* ribosomal proteins S8 and L6. Nucleic Acids Res. 12:8209-8217.

111. Muto A., Osawa S. (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. USA 84:166-169.

112. NAG Library Services Manager, NAG Central Office, Mayfield House, 256 Banbury Road, Oxford OX2 7DE.

113. Nei M., Tajima F. (1981) DNA polymorphism detectable by restriction endonucleases. Genetics 97:145-163.

114. Nevins J.R. (1983) The pathway of eukaryotic mRNA formation. Ann.Rev.Biochem. 52:441-466.

115. Newgard C.B., Nakano K., Hwang P.K., Fletterick R.J. (1986) Sequence analysis of the cDNA encoding human liver glycogen phosphorylase reveals tissue-specific codon usage. Proc. Natl. Acad. Sci. USA. 83:8132-8136.

116. Nishimura S. (1978) Modified nucleosides and isoaccepting tRNA. In: Transfer RNA pp168-19. (Altmann S., ed) MIT press, Cambridge, Massachusetts.

117. Nishisato S. (1980) Analysis of categorical data: Dual scaling and its applications. University of Toronto Press, Toronto.

118. Nomura M., Post L.E., Jinks C.S. (1980) New aspects of regulation of ribosomal protein genes in *Escherichia coli*. In: Genetics and evolution of RNA-Polymerase, tRNA and Ribosomes. (S. Osawa, ed.), University of Tokyo Press, Tokyo, and Elsevier/North-Holland, pp315-328.

119. Nussinov R. (1981) The universal dinucleotide asymmetry rules in DNA and the amino-acid codon choice. J.Mol.Evol. 17:237-244.

120. Ogasawara N. (1985) Markedly unbiased codon usage in *Bacillus subtilis*. Gene 40:145-150.

121. Palmer J.D. (1985) Evolution of chloroplast and mitochondrial DNA in plants and algae. In: MacIntyre R.J. (ed) Molecular evolutionary genetics. Plenum, New York, pp131-240.

122. Pearson K. (1901) On lines and planes of closest fit to a system of points in space. Philosophical Magazine and Journal of Science, Series 6. 2:559-572.

123. Pfitzinger H., Guillemaut P., Weil J-H., Pillay D.T.N. (1987) Adjustment of the tRNA population to the codon usage in

chloroplasts. Nucleic Acids Res. 15:1377-1386.

124. Pieczenik G. (1980) Predicting coding function from nucleotide sequence or survival of 'fitness' of tRNA. Proc. Natl. Acad. Sci. USA 77:3539-3543.

125. Ponnuswamy P.K., Muthusany R., Manavalan P. (1982) Amino acid composition and thermal stability of proteins. Int.J.Biol.Macromol. 4:186-190.

126. Preer J.R., Preer L.B., Rudman B.M., Barnett A.J. (1985) Deviation from the universal code shown by the gene for surface protein 51A in *Paramecium*. Nature 314:188-192.

127. Rennell D, Poteete A.R. (1985) Phage 22 lysis genes: Nucleotide sequences and functional relational with T4 and λ genes. Virology 143:280-289.

128. Resnick M.A. (1970) Sunlight-induced killing in *Saccharomyces cerevisiae*. Nature 226:377-378.

129. Rosenberg A.H., Simon M.N., Studier F.W. (1979) Survey and mapping of restriction cleavage sites in bacteriophage T7 DNA. J.Mol.Biol. 135:907-915.

130. Rowe G.W., Szabo V.L., Trainor L.E.H. (1984) Cluster analysis of genes in codon space. J.Mol.Evol. 20:167-174.

131. Rubin G.M. (1983) Dispersed repetitive DNAs in *Drosophila melanogaster*. In: Mobile genetic elements (Shapiro J.A., ed), Academic Press, London.

132. Salinas J., Zerial M., Filipski J., Bernardi G. (1986) Gene distribution and nucleotide sequence organization in the mouse genome. Eur. J. Biochem. 160:469-478.

133. Sharp P.M., Rogers M.S., McConnell D.J. (1985) Selection pressures on codon usage in the complete genome of bacteriophage T7. J.Mol.Evol. 21:150-160.

134. Sharp P.M., Tuohy T.M.F., Mosurski K.R. (1986) Codon usage in yeast: Cluster analysis clearly differentiates between highly and lowly expressed genes. Nucleic Acids Res. 14:5125-5143.

135. Sharp P.M., Li W-H. (1986) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. Nucleic Acids Res. 14:7737-7749.

136. Sharp P.M., Li W-H. (1987) An evolutionary perspective on synonymous codon usage in unicellular organisms. J.Mol.Evol. 24:28-38.

137. Shepherd J.C.W. (1981) Method to determine the reading

frame of protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc. Natl. Acad. Sci. USA 78:1596-1600.

138. Shields D.C., Sharp P.M. (1987) Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. Nucleic Acids Res. 15:8023-8040.

139. Singer C.E., Ames B.N. (1970) Sunlight ultraviolet and bacterial DNA base ratios. Science 170:822-826.

140. Singhal R.P., Fallis P.A.M. (1979) Structure, function, and evolution of transfer RNAs (with appendix giving complete sequences of 178 tRNAs). Prog. Nuc. Acid. Res. Mol. Biol. 23:227-263.

141. Sjostrom M., Wold S. (1986) A multivariate study of the relationship between the genetic code and the physical-chemical properties of amino acids. J.Mol.Evol. 22:272-277.

142. Staden R. (1984) Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. Nucleic Acids Res. 12:551-567.

143. Starmer W.T., Ganter H.J. (1986) Quantum and continuous evolution of DNA base composition in the yeast genus *Pichia* Evolution 40:1263-1274.

144. Storck R., Alexopoulos C.J. (1970) Deoxyribonucleic Acid in fungi. Bacteriol. Rev. 34:126-154.

145. Sueoka N. (1961a) Variation and heterogeneity of base composition of deoxyribonucleic acids: A compilation of old and new data. J.Mol.Biol. 3:31-40.

146. Sueoka N. (1961b) Compositional correlation between deoxyribonucleic acid and protein. Proc.Nat.Acad.Sci. USA 26:45-52.

147. Sueoka N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. USA 48:582-592.

148. Sueoka N. (1986) Directional mutation pressure and dynamics of DNA sequence. Jpn.J.Genet. 61:636-643.

149. Treffers H.P., Spinelli V., Belser N.O. (1954) A factor (or mutator gene) affecting mutation rates in *E.coli* Proc. Natl. Acad. Sci. USA 40:1064-1071.

150. Wada A., Suyama A. (1985) Third letters in codons counterbalance the G+C content of their first and second letters. FEBS letters 188:291-294.

151. Weir B.S. (1985) Statistical analysis of molecular genetic data. I.M.A. J. Math. Applied. Med. & Biology 2:1–39.

152. Yamagishi H. (1974) Nucleotide distribution in bacterial DNAs differing in G+C Content. J.Mol.Evol. 3:239–242.

153. Yamao F., Muto A., Kawauchi Y., Iwami M., Iwagami S., Azumi Y. and Osawa S. (1985) UGA is read as tryptophan in *Mycoplasma capricolum* Proc. Nat. Acad. Sci. USA 82:2306–2309.